THE PREVALENCE OF CORRESPONDENCE COMPUTATIONS IN VISUAL PROCESSING

by
Zheng Ma

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March, 2016

Abstract

Correspondence computation is the general process during which received perceptual signals are assigned to the same or different sources. It is a pervasive process involved in many visual cognitive tasks. People need to make this computation both for simultaneously received signals and for temporally separated signals. Due to the noisy nature of visual input signals, it is an error-prone process. Therefore, it serves as the cause of many limits in human performance. However, its role in human cognitive abilities has often been ignored or underestimated. Besides, how correspondence computations are done, and how it is related to other computations in the visual system, haven't been explored a lot. In this dissertation, I combined evidence from human behavioral experiment, computational modeling, and testing of brain-damaged patient to investigate these questions.

In Chapter 2, I showed that human participants couldn't accurately report the number of presented objects in a typical visual working memory task. I then showed that a clustering algorithm with noise in the correspondence process could simulate human performance very well. Therefore, parts of the limits in human memory ability could be explained by imperfect correspondence computations.

In Chapter 3, I explored how different correspondence computation algorithms are combined to solve the problem of motion direction judgment. I found that a lower-level luminance transient detection system and a higher-level position comparison system work complementary to each other. The relative contribution of each system depends on the signal strength of that system.

In Chapter 4, I showed that limit in object tracking ability is a result of noisy visual inputs, suboptimal eye-movement strategy, and probabilistic correspondence computations. No external resource-like limits are needed to understand human's limited capacity in tracking multiple objects at the same time.

In sum, these results suggest that correspondence computations play important roles in visual cognition. It is pervasive, and a failure in this process could lead to further failures in related tasks. Human visual cognition should be understood in terms of the involved computations and the possible errors could arise during these computations.

Readers:

Jonathan Flombaum (advisor)

Shreesh Mysore (second reader)

Howard Egeth

Jason Fischer

Michael McCloskey

Colin Wilson

Kyle Rawlins

Acknowledgments

Five years ago when I got an offer from Dr. Jonathan Flombaum, I couldn't imagine I can really get a PhD degree in cognitive psychology. I believe it's all of the support and help I got from professors, colleagues, friends and family members that made my dream come true.

I would like to thank Jon for all of his support during the past five years. I could not imagine getting my PhD without Jon being my advisor. Jon helped me grow up both as a researcher and as a human being. He is both smart and hardworking. He is always very patient to me. I really hope one day I will become a researcher just like him. I also want to thank Jon's wife Jojo and his three kids. Thank you all for making my life so colorful!

I would also want to thank all of the professors I have worked with, and I am so lucky that some of them have become my committee members. To Dr. Michael McCloskey, I have learned so much from you. You are always serious about every research question. You made me know that we need to make careful logical inference before starting doing any experiments. To Dr. Colin Wilson, I am so grateful to have been in your Bayesian inference class, and to have collaborated with you. You are definitely my math and cognitive science hero! To Dr. Howard Egeth, I think I should have talked to you more often. I benefited so much from each conversation with you. Thank you for all of your help in experimental design and paper recommendations. I really hope I can continue to hear you talking about 'a 1970 paper'. To Dr. Shreesh Mysore, though I was the TA of your class, I felt I've learned more than I've contributed. Thank you so much for helping me with different stats problems. To Dr. Jason Fischer,

thank you so much for being my committee member! I haven't talked to you very much, but your questions during my oral defense are both hard and inspiring.

Besides my committee members, there are many other professors that have helped me a lot through my five-year graduate school life. To Dr. Justin Halberda, I appreciate every single thing you have taught me! I miss those times when we still have lab meetings and we can discuss about interesting scientific questions together. I hope I can be as energetic as you when I am in your age! To Dr. Lisa Feigenson, thank you for all of the encouragements you gave me. I still remember during my advanced exam you asked me a lot of 'hard' questions. And those helped me gradually know what I really want from my studies. To Dr. Marina Bedny, I really appreciate all of our conversations during and after my PBS talks. They help me a lot in both organizing talks and in thinking about research questions. I also want to say many thanks to Dr. Steven Yantis. I know Steve's name when I was an undergraduate students and reading famous papers in visual cognition. I was so lucky to have been in the same department with him. I will always miss him, and try my best to become some one like him.

I couldn't finish my PhD degree without the support from my colleagues and friends. To Gi-Yeul Bae, I always feel you are my big brother. In my mind, you are my real 'oppa'! Thank you for sharing your thoughts and data with me. I hope I can join you in California very soon. To Hee-Yeon Im, if Gi-Yeul was my brother, then you are my sister. I miss so much the time when our desks were facing to each other. To Shenghua Zhong, I think I owe you too much. You taught be modeling, you taught me computer science, and you even found me a husband! I hope we can have more and more collaborations in the future! To Mark Schurgin and Feitong Yang, I am so grateful to

have you as my lab members. Whenever I had problems in my research, you are always there to help me. To Kitty Zhe Xu, we are almost done! I miss the time when we laugh together and cry together. Let's be together again in the Bay area! To Jenny Wang, I can't imagine how my PhD life would be like without you. It's so great that I can have someone like you to discuss almost everything, everyday, together! To Robert Eisinger, Tyler Knowlton, Xiaomo Chen, Yi-shin Sheu, Li Guo, Iven Yu, and my dear roommate Yang Dong, sometimes when I was sad, even just talking to you would make my day better!

Finally, I would like to thank my family members. To my parents and grandparents, I'm going to acknowledge them in Chinese. 谢谢爸爸妈妈姥姥姥爷在过去二十多年里给予我的爱和支持。在异国他乡读书虽然辛苦，但是只要想到你们我就觉得自己还是一个幸福的无忧无虑的小孩。我爱你们，也希望你们能一直健康开心下去！To my husband Jianqiao Feng, I am so grateful to meet and get married with you. And I think I would be luckier if I could know you earlier when I was in Beijing. I always feel that even without getting my PhD, my time at Hopkins will not be a waste of time as long as I can still know you and be your wife.

Thank you all! Thank you PBS and Hopkins! I will try to be a better person, and I will always miss the past five years.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 The prevalence of correspondence computations

The visual system is often understood as an information processing system. But compared to systems engineered by people, visual processing is unusual in that the sender of the signals, the visual world, does not have an explicit goal to communicate with the receiver. The result is that in addition to making inference about the content of the signal, the visual system also needs to determine, on the basis of the received signals, what and how many sources are present. Broadly, I will refer to the assignment of signals to sources as 'correspondence'.

Correspondence computations must take place pervasively in visual processing. However, since these computations could often be solved without awareness, one does not always think about the involvement of correspondence computations in visual cognition. For example, when pictures of two objects are superimposed on each other (**Figure 1.1**), one can easily recognize both of them and selectively pay attention to either one (Goldstein & Fink, 1981). In this case, a correspondence computation is needed to assign simultaneously received signals to different sources. Take another example, when one sees an object moving along a trajectory, he or she will not think there are different objects flashing at different locations. In this case, a correspondence computation is needed to assign temporally separated signals to the same source. These two examples suggest that both signals received simultaneously and signals received across time need to be assigned to their sources correctly. Note that these two types of correspondence computations are unlikely to rely on totally different mechanisms. However, this distinction is helpful for descriptive reasons.

Figure 1.1 A superimposed image of a face and a house (adopted from Kanwisher & Wojciulik, 2000). One easily recognizes both objects easily. Though human observers are often unaware of it, a kind of correspondence computation is needed to assign overlapped signals to different sources.

Before moving to more specific discussions on the role of correspondence computations, it is worth discussing the relationship between correspondence computation and a very similar term, the 'binding problem'. Under Treisman's definition (Treisman, 1995), the general binding problem is very similar to the correspondence computation discussed here. Treisman explicitly stated that 'Some mechanism is needed to "bind" the information relation to each object and to distinguish it from others'. However, by using the term 'binding problem', Treisman and other researchers treated the process more like a consequence of some other computations (e.g. attention), rather than a computation that is the basis of many visual cognitive processes. Since the goal of the current dissertation is to study how correspondence computation is involved and completed in different cognitive tasks, I will stick to the term 'correspondence computation' to emphasize the computational nature of the process. I will use the term

'binding problem' to refer to some specific cognitive processes that correspondence computation is involved.

In the next section, I will first give a brief summary of the prevalence of the two kinds of correspondence computations: assigning simultaneous signals and temporal separated signals. I will show that many classic visual phenomena could be understood under the framework of correspondence computations. This perspective will provide a unified framework to explain many problems.

1.1.1 Corresponding simultaneously received visual signals to different sources

If we only look at the signals on the retina, there is no direct information of which signals are from the same source. This structure is also mirrored in the primary visual cortex, which is spatially organized. Since most objects often occupy the receptive field of multiple primary visual cortex neurons, at this level it's not clear which neurons are firing for the same object. These problems are finally revealed at higher-level visual cognitive tasks. To detect and recognize objects in a visual scene, a massive amount of correspondence computation is needed to calculate the relationship among multiple signals.

Many classic visual phenomena in fact involve this kind of correspondence computation, but have not been discussed explicitly in this way. For example, it has been well studied that when an object is presented with nearby flankers in the visual periphery, visual crowding will happen such that one can hardly recognize the properties of the object (Intrilligator & Cavanagh, 2001). Moreover, one sometimes reports the crowded object as taking characteristics of the nearby flankers (Whitney & Levi, 2011). This crowding effect in fact reveals a failure of the correspondence computation, during which

the signals were assigned to wrong sources. Therefore, crowding could be understood as the difficulty of correspondence computations in the periphery.

How people can bind different features together is also a type of correspondence computations. It has been shown that when stimuli were presented with many distractors in the peripheral visual field, it's very hard for normal observers to bind color and orientation information correctly (Neri & Levi, 2006). Similar difficulties have also been observed with brain-damaged patients. Friedman-Hill, Robertson, and Treisman (1995) reported a case study of a patient, R.M., who has damage in bilateral parietal-occipital areas. R.M. have deficit in binding features of the same object together. For example, when presented with a red O and a green X, R.M. might report seeing a red X. In one way, this feature-binding problem could be treated as an independent cognitive process. In the other way, it could be treated as an example of the general correspondence problem, during which different properties of the same source need to be integrated together. The second approach provides a more unified answer to different phenomena observed together with the feature-binding problem. For example, the patient R.M. also shows deficits in searching for targets defined by multiple feature dimensions (conjunction search), and in counting the number of objects (Robertson et al., 1997). Under the correspondence computation framework, these two deficits are not surprising because they all involve assigning multiple simultaneously received signals to different sources. In Chapter 2, I will discuss more about the relationship between number estimation and the correspondence computation.

A final example of simultaneous correspondence computation comes from the object-based attention literature. It has been shown that things presented on the same

object can share some advantages in visual processing (Egly, Driver, & Rafal, 1994; Scholl, 2001). One pre-assumption of this advantage is that people can successfully construct representations of individuated objects from noisy input signals. However, 'objects' do not automatically arise from the signals at the retina or at the primary visual cortex. To selectively attend to the human face in **Figure 1.1**, one first needs to decide which signals come from the face. Therefore, correspondence computation plays a fundamental role in object-based attention.

1.1.2 Corresponding visual signals received across time

Signals need to be assigned to sources not only simultaneously, but also across time. Again, the correspondence computation of temporally separated signals is pervasive in visual cognition, but not often understood in this way.

Motion perception is the simplest example that involves correspondence computation across time. Signals received at different time points need to be assigned correctly to form a coherent representation of directional movement. More importantly, motion perception is special because at least two systems can complete the correspondence computation: a lower-level motion energy system that can detect simple luminance change, and a higher-level system that requires the comparison between object positions (Lu & Sperling, 1996, 2001). The representation of the lower-level system allows the correspondence computation to be done implicitly. Neurons in the lower-level system are sensitive to luminance change within a certain spatial-temporal change. If two signals happen within one neuron's receptive field, the neuron will fire, and thus the two signals are assigned to the same source automatically. In other words, the lower-level system directly represents correspondence relationships. On the other hand, the higher-

level system first represents the signals' variable values (e.g. time and location), and then makes explicit comparison of these values to perform the correspondence computation. Despite the difference in the representation and algorithms used by the two systems, the outputs of both systems are matched motion representations across time. In Chapter 3, I will closely explore the relative contribution of the two systems to motion perception.

Tracking correspondence across time is not only important for motion perception per se, but also important for stable representation for the same object across time. The same object can look very different under different illumination, orientation, and occlusion conditions (DiCarlo & Cox, 2007). It is very important for the visual system to form a stable representation of the same object. Correspondence computation is needed here to assign slightly different signals to the same object identity. It has been proposed that both feature stability and spatiotemporal continuity are used to make this kind of correspondence computation (Yi et al., 2008). Feature stability refers to the relative consistent appearance of the same object under different conditions. Spatiotemporal continuity refers to the fact that objects either stay at the same spatial location, or move along continuous (or approximately continuous) trajectories, but are never able to discretely jump between locations that are very far from each other. Therefore, signals share similar feature properties and obey spatiotemporal continuity are often integrated together to form coherent object representations. No matter which rule is applied, the take home message is that correspondence computation is very important for stable object representation.

1.2 Correspondence computation at the algorithmic level

Correspondence computation is pervasive in visual processing, but why sometimes it's very hard to get the correct correspondence relationship? And how it could be done at the algorithmic level? These are two closely related questions, and in fact, the difficult of correspondence computation need to be understood at the algorithmic level.

At least two reasons can cause failures in the correspondence computation process. The first reason is related to the input signals of correspondence computation. The visual system is built in a way that the visual field, especially the peripheral visual field, can only be represented noisily (Carrasco & Frieder, 1997). Therefore, for most of the areas in the visual field, one can only rely on noisy inputs to make inference on correspondence relationships. Therefore, it's very challenging for any correspondence algorithm to recover the accurate source information.

The second reason is that multiple systems/algorithms could be used to compute correspondence relationships for the same inputs. The visual system needs a way to integrate multiple outputs to form a single perception. Although in most cases, different systems and algorithms are signaling the same correspondence relationship, in some special cases, different algorithms can actually lead to conflict outputs. For correspondence of simultaneously received signal, it has been proposed that observers use spatial proximity to bind features and parts together into an object (Treisman & Gelade, 1980). However, there can be different levels of spatial proximity. In a famous painting by Salvador Dali (**Figure 1.2**), it is unclear whether the two areas highlighted by the two green boxes came from the same source: the face of a gentle man, or came from

two separate sources: a farmer and a young lady. In this case, spatial proximity at local areas and global areas are conflicting with each other, and that's why observers are uncertain about the correspondence relationship and the image becomes bi-stable.

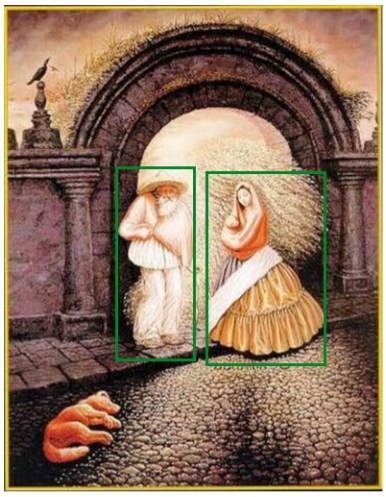


Figure 1.2 A painting by Salvador Dali. The area in the two green boxes could be either interpreted as the parts of a larger object, or as two individual objects by themselves.

Different algorithms of the temporally separated correspondence computation can also signal opposite outputs. Dawson (1991) has proposed three principles for the higher-level motion system to make correspondence judgments: a nearest neighbor principle (the system prefers short displacement to long displacement), a relative velocity principle (the system prefers neighboring element to have identical velocities), and an element integrity principle (the system penalizes splitting or merging events). A conflict between the nearest neighbor principle and the relative velocity principle could be easily seen in a classic apparent motion display, the Ternus display (**Figure 1.3**). In a typical Ternus display, a set of equally distributed objects was first presented to the observer (at position 1, 2 and 3). Then after a variable blank interval, another set of objects was presented, with a displacement by the distance between neighboring elements (at position 2, 3 and

4). When the two frames of objects were presented alternatively, two possible types of motion could be perceived (Petersik & Rice, 2006). If the nearest neighbor rule is applied, then objects at position 2 and 3 in the first frame should be assigned to objects at position 2 and 3 in the second frame, so that zero displacement has happened to each of them. In this case, the leftmost object in the first frame and the rightmost object in the second frame have to be assigned to the same source. This algorithm will produce 'element motion perception', in which the middle objects are stationary and the outside object is moving back and forth between the two outer positions. If the relative velocity rule is applied, then objects at position 1, 2, 3 in the first frame should be assigned to objects at position 2, 3, 4 in the second frame respectively, such that all objects are moving at the same velocity. This will produce 'group motion perception', in which the objects seem to form a group that moves back and forth together.

In reality, human observers perceive both element motion and group motion, but under different interstimulus interval (ISI) conditions. Shorter ISIs between the two frames often make people see more element motion, and longer ISIs often make people see more group motion. It's still an open question that how different algorithms are interaction with each other to generate the final perception. In Chapter 3, I will try to use an empirical experiment to answer this question.

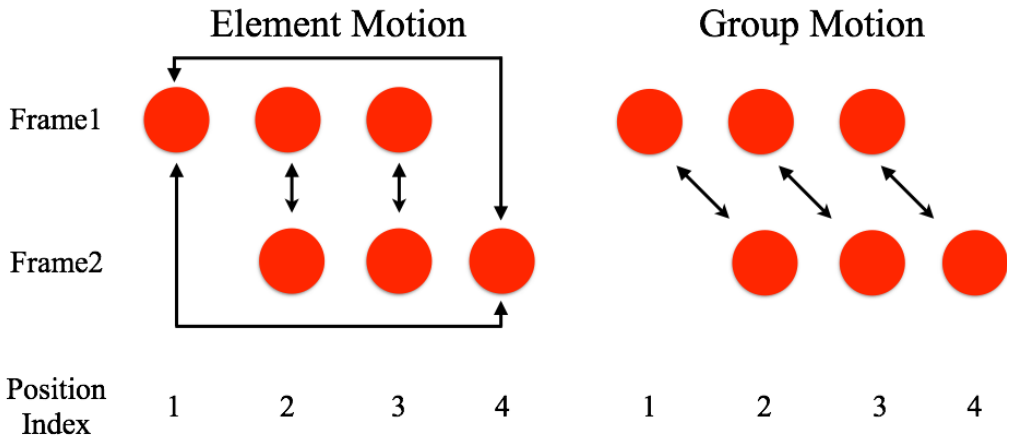


Figure 1.3 An illustration of the Ternus display. With the same visual inputs, two correspondence relationship could be calculated, and are related to two different perceptions.

In sum, correspondence computation is not easy because 1) it often relies on noisy input signals, 2) different correspondence algorithm can generate different outputs. In Chapter 2 and 4, I will explore how noisy inputs and some form of correspondence computation can lead to observed limits in visual cognition. In Chapter 4, I will take motion perception as a special example to study the relative contribution of different algorithms to compute correspondence relationship

1.3 Correspondence computation as a primary constrain on human visual cognition

The current dissertation will elaborate specific application of correspondence computation in a broad range of visual cognitive processes. Emphasizing the role of correspondence is important not only because it can characterize computations involved in vision, but also because it can explain many limits observed in human behavior. Correspondence computations are not always perfect, and thus can lead observers to make errors. If one does not look at the algorithmic level of the problem, one may mistake these errors for other reasons. For example, a car cannot run faster than 300

miles per hour. If one does not look at the specific designs and mechanisms of the car, one may think this is due to the car doesn't have enough gas. However, the true reason is obvious when one knows the principles of how cars work: inherent limits in the design of the car and constraints in the environment make it impossible for the car to run that fast. Similarly, understanding how computations are done in the human visual system will help us fully understand the limits in human behavior.

Next, I'm going to give a brief summary of some representative cognitive processes that clearly involve correspondence computations. For each process, I will first give a short review on how human limits in that task are often explained. Then, I will show why an explanation at the correspondence computation level is a better way to help us understand human visual cognition.

Multiple Object Tracking

Tracking moving objects across time is probably the most straightforward task to study how correspondence computations are used in visual cognition. One can clearly feel the correspondence computations involved in this task. In order to make sure the same object is tracked correctly, the observer needs to constantly assign the signals received at the current moment to those received before (Scholl, 2009). More importantly, it has been widely shown that human tracking ability is extremely limited. When participants are asked to track a set of moving targets among a larger set of featurally indistinct objects during a multiple object tracking (MOT, Pylyshyn & Storm, 1988) task, they can only track three to five targets successfully. Longer duration, faster moving speed, and a decrease in object proximity also lead to significant impairment in tracking performance (Pylyshyn, 2004; Alvarez & Franconeri, 2007). Therefore, object

tracking also provides a great opportunity to study how correspondence computation is related to limits observed in human behavior.

Despite the clear involvement of correspondence computation during MOT, its role has been largely underestimate in previous literature. The limits in the tracking task have often been explained by a form of limited available cognitive resources, with either a fixed-slots or flexible nature (e.g. Cavanagh & Alvarez, 2005; Drew & Vogel, 2008; Pylyshyn, 2001; Alvarez & Franconeri, 2007; Horowitz & Cohen, 2010; a & Huang, 2009; Holcombe & Chen, 2011). Computations are not explicitly mentioned in these theories.

This kind of resource-based explanation has many problems. First, it is unclear what the resource is, where it comes from, and how it is represented in the human brain. In the current articulation of these theories, the word "resource" is more like a replacement of tracking limits. And thus, sometimes it feels like circular reasoning: since the cognitive resource is discrete/flexible, the observed tracking precision is discrete/flexible, and thus the cognitive resource is discrete/flexible. This kind of reasoning is similar to the claim that a homunculus is inside our brain and controlling our mind, and doesn't add much to our knowledge of why human cognitive abilities are limited.

Second, the role of correspondence computations in visual tracking is ignored. With a special emphasis on resource, many current theories assume the correspondence computation among simultaneously received signals could be done perfectly (e.g. Ma & Huang, 2009; Vul et al., 2009). They also do not have explanations on how target identities are corresponded across different time frames during tracking. Therefore, errors

arise during the correspondence computations may have been mistakenly interpreted as evidence for resource limits.

There is some evidence making me believe that correspondence computation can explain performance limits in MOT. In one previous study, I have shown that people are uncertain about the number of presented object even at the beginning of an MOT trial, probably due to a failure in simultaneous correspondence computation (Ma & Flombaum, 2013). Besides, preventing corresponding errors during MOT could significantly improve tracking performance (Bae & Flombaum, 2012). In Chapter 4, I will show how probabilistic correspondence computations based on noisy input signals could lead to limits in human tracking performance that are often explained by limited resources.

Visual Working Memory

Visual working memory (VWM), especially the kind of task that is used to study the VWM ability, is another example that correspondence computation is needed. More specifically, two paradigms have been developed to study VWM ability. In the change detection paradigm, participants are asked to determine whether two sequentially presented frames contain the same or different stimuli (Luck & Vogel, 1997). In the delayed estimation paradigm, participants are asked to remember a set of stimuli, and then report the property of a probe item from many provided choices (Zhang & Luck, 2008). In both paradigms, participants need to give answers about specific objects/items: whether an item has changed or what is the property value of an item. To do the tasks, it's critical for participants to correctly assign noisy signals to their sources, and rely on the right representation of the probed location.

The role of correspondence computation has also been underestimated to explain

limits in VWM ability. Although within a really brief memory duration (<250 ms), people can remember up to eleven items (Sperling, 1960), for longer durations (e.g. 900 ms), people can only remember up to four objects with high precision (e.g. Luck & Vogel, 1997; Bays & Husain, 2008). To explained this limited memory ability, both fixed-slots and flexible resource theories have been proposed to explain the observed limits in human performance (e.g. Zhang & Luck, 2008; Luck & Vogel, 2013; Alverez & Cavanagh, 2004; Anderson, Vogel, & Awh, 2011; Bays & Husain, 2008; Wilken & Ma, 2004). Almost none of these theories consider how computations are done to form representation of objects during VWM tasks. They also often assume that signals can always be assigned to the correct source, and ignore the possibility of correspondence errors. However, it has been shown that participants sometimes report the property of a nontarget item (Bays, Catalao, & Husain). Taking the advantage of integral features to prevent correspondence computation errors could also improve memory performance (Bae & Flombaum, 2013). Therefore, at least some of the VWM limits can be explained by errors in correspondence computation. In Chapter 2, I will explore how simultaneous correspondence computation at the beginning of VWM tasks could lead to errors in memory responses.

Visual Crowding

As discussed in section 1.1.1, representing integrated object representation requires correspondence computation to be done correctly. One can clearly feel the failure of simultaneous correspondence computation in the crowding effect, during which nearby flanker objects impair human observer's ability to discriminate and individuate a peripheral presented object (Intrilligator & Cavanagh, 2001). In other words, a critical

spacing between the target object and its neighbors is needed for an observer to recognize the target (Pelli & Tillman, 2008). If the spacing is not large enough, signals from the target and flanker objects cannot be assigned correctly to their original sources.

Unlike the MOT and VWM literature, resource-based theories haven't played a dominant role in the explanations of the crowding effect. In fact, one can easily come up with a 'flexible resource' theory of crowding. For example, crowding could be explained by saying cognitive resources is less available in the peripheral so that when more than one object are presented, the recognition ability is low. However, similar to the shortcomings of the other resource theories, this kind of explanation does not provide any more knowledge than just describe the phenomenon.

Probably this is why many of the current theories in the crowding literature tried to explain the phenomenon in computational terms. For example, Wilkinson et al. (1997) proposed that crowding is the result of mutual inhibition among complex cells and simple cells. van den Berg et al. (2010) suggested that population coding principle could account for many phenomena observed in crowding. Dayan and Solomon (2010) utilized a Bayesian inference approach and claimed that crowding is the consequence of making inference based on information collected from a relatively larger receptive field in the periphery. Similarly, Balas, Nakano, and Rosenholtz (2009) proposed that information in the periphery is represented by summary statistics, and thus target and flanker information will be mingled together under crowding condition. These theories, no matter based on neuronal data or modeling data, do care about the representations and computations in the brain. They provided some possibilities that how a failure in signal-source correspondence could happen, and how it could further limit visual processing. In

Chapter 3, I will use crowding as a tool to study how different correspondence algorithms are involved in motion perception.

1.4 Outline of the current dissertation

In this dissertation, I will propose a framework of understanding human visual cognition based on correspondence computations. I will use visual working memory, motion perception, and object tracking as three representative examples to show that correspondence problems are affecting different visual processing in similar ways.

In Chapter 2, I will explore the correspondence computation happens at the beginning of a typical VWM task. I will first use behavioral data to suggest people are not good at individuating objects. Then, I will propose a computation model that considers noisy correspondence computation and can simulate human performance.

In Chapter 3, I'm going to use motion direction judgment as a case study to explore how different correspondence algorithms are used for perceptual judgments. More specifically, I want to know how the higher-level position comparison correspondence process would be used together with lower-level luminance detection system.

In Chapter 4, I'm going to ask how correspondence analysis can support higher-level activity of object tracking. With the physical constraints of the human visual system, how much variance in human performance could be explained by correspondence computation? With both behavioral data and model simulation, I'm going to suggest that after considering perceptual limits, eye-movements, and correspondence computations, no external resource constraint is needed to explain human tracking limits.

Chapter 2: Individuating objects in visual working memory as a correspondence problem

Visual working memory (VWM) plays an essential role in human daily life. Many studies have talked about the content of human VWM in terms of items or objects. For example, Zhang and Luck (2008) claimed that "human observers store a high-resolution representation of a subset of the *objects*". It's clear that many studies have discussed VWM representations in terms of *objects*, but did not provide detailed explanations on how these representations of objects were formed (e.g. Zhang & Luck, 2008; van den Berg et al., 2012). They didn't talk about what kind of computation can lead to these memory contents, and how the errors in these computations can lead to response errors in the memory task.

At the very beginning of the memory process, one does not directly represent the objects presented on the display. This is because human eyes don't have sensors that can directly detect objects. Observers only have access to light signals projected to our retina. What makes the problem complex is that one often gets multiple signals from a single object. What's more, there is often a lot of noise in these signals. Therefore, at least at early stages, memory representations can be thought as noisy signals generated from the objects. Then, in order to form representations of objects, a type of correspondence computation is needed to assign these signals to different sources.

Any failure in this signal-source assignment process could result in inaccurate object individuation: multiple signals from an object may be misinterpreted as from different objects, and signals from different objects may be misinterpreted as from the

same objects. Therefore, there may not be a one-to-one accurate correspondence relationship between real presented objects and memory-represented objects.

Previous research on object individuation

In a previous study (Ma & Flombaum, 2013), I studied whether human participants have perfect object individuation ability at the beginning of MOT tasks. The logic was that if one can accurately individuate objects and represent the presence of each of them, one would be able to report the number of objects correctly. However, if the observer doesn't know the number of objects, it could be the consequence of noisy correspondence algorithms. Surprisingly, this kind of data hasn't been collected before this study.

Figure 2.1 plots the result I obtained from the first experiment of the study. In that experiment, participants were asked to remember four to nine target locations out of twenty four object locations for 0.5 seconds. Then they were asked to click all target locations. I unconstrained the number of responses they could make, such that they were allowed to click as many objects as they thought were targets. The number of selected objects then could serve as a measure of participants' knowledge of target number. In **Figure 2.1**, the x-axis indicates the true number of targets, and the y-axis indicates the number of responses participant made in one trial. The darkness of each cell reflects the proportion of responses. The results showed that at the beginning of an MOT trial, people are uncertain about how many objects they were asked to track, especially for trials with higher target loads. It's clear in **Figure 2.1** that people's number responses were very noisy, and both under- and over-estimation happened a lot. This suggested that people cannot always accurately individuate objects, and this is probably due to errors in

correspondence computations of simultaneously received signals. Furthermore, this uncertainty in input at the beginning of trial may further lead to errors during tracking.

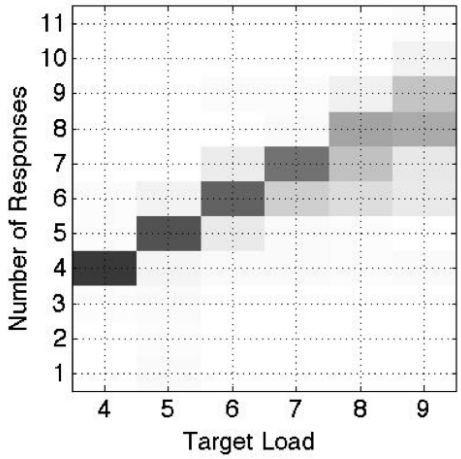


Figure 2.1 Result of the first experiment in Ma and Flombaum (2013). Only result from the 0.5s condition is plotted here. The saturation of the cells are in proportion to the count of a given number response.

In this chapter, I tried to extend these findings to object representations in VWM. More specifically, I was interested in how human observers are representing signals in VWM, and how correspondence computation could lead to inaccurate object individuation.

I first used a behavioral experiment to get an empirical measure of people's enumeration ability in a typical VWM task display. The result served as baseline evidence that people do not have perfect object individuation ability. Then, I used a clustering algorithm to investigate how noisy correspondence computation could lead to inaccurate object individuations.

2.1 Experiment 1: A behavioral measure of inaccurate object individuation in human VWM

2.1.1 Method

Participants

24 Johns Hopkins University undergraduates participated in this experiment. All had normal or corrected-to-normal visual acuity. The protocols of this and all the reported experiments were approved by the Homewood Institutional Review Board of Johns Hopkins University.

Apparatus

The experiment was done in a dark room with the computer monitor as the only light source. All stimuli were presented on a calibrated CRT monitor at a viewing distance of 60 cm such that the whole display subtended approximately 31.7° by 24.5° degree of visual angle. The area used to present stimuli was about 12.5° × 9.4°. The refresh rate was 120 Hz.

Stimuli and Procedure

We used the same 180 equally spaced colors that were used in Bae, Olkkonen, Allred, & Flombaum (2015). These colors only varied in hue in CIELAB space (L*=70, a*=0, b*=0, radius of 38), and are all within the monitor gamut. We used the color of the center point of the chosen CIELAB hue ring as the background color.

In each trial, one to eight color squares could appear on the screen. We randomly generated 320 trials, 40 trials for each memory load condition. The colors and locations of the squares were randomly selected, with the constraints that 1) Each square subtended 0.94° × 0.94° and was at least 2° (measured border to border) away from each other; 2)

Each square took one of the 180 possible colors, and the distance in the color space between any two squares could not be smaller than 10 colors.

At the beginning of each trial, a 0.6° × 0.6° white fixation cross was presented for 500ms. After a 500ms blank duration, the memory display (one to eight colored squares) was presented for 100ms. After another 900ms blank delay period, one of three testing conditions could happen. In the Color condition (standard delayed estimation task), all of the original squares were presented, but only with white outlines and no color information was provided. One of the squares was cued with a thicker outline, suggesting that it was the target square that the participant needed to make response for. A response color wheel and a mouse cursor were presented together with the square outlines. The color wheel contained all 180 colors that could have been presented. The participants were asked to click at the specific color that was presented in the target square, as accurate as possible (see **Figure 2.2** for an illustration of experimental procedure).

Figure 2.2 Procedure of the working memory tasks.

In the Number condition, the following sentence appeared on the screen: “Please type in the number of presented squares”. No other information was given, and the participants were asked to type in the number of squares they thought were presented, as accurately as possible.

In the Location condition, the following sentence, together with a mouse cursor, appeared on the screen: “Please click at all of the object locations”. The participants were asked to click all locations where they saw an object. After they thought they had clicked all locations, they clicked at an ‘OK’ button to complete the trial.

All participants completed the same 320 trials with the same object selected as the target. However, different trials were assigned to different testing conditions for different participants. The assignments of trials and participants were randomly counterbalanced with the following constraints: 1) Each participant completed all 320 trials, 40 trials for each memory load condition; 2) For each participant, among the 40 trials of each memory load condition, 50% (20 trials) were the Color response condition, 25% (10 trials) were the Number response condition, and the remaining 25% (10 trials) were the Location response condition; 3) For each specific trial, it was tested in the Color condition by half of the participants (12 participants), and was tested in the Number and Location condition by 25% of the participants (6 participants) respectively.

The Color condition was mainly designed to make sure the participants were doing the color VWM task, not only counting or remembering the location of the squares. Moreover, the Color condition was the major task for all participants, making them to treat the memory task as their primary concern. The Number and Location conditions supplied two different measurement of participants’ number knowledge of the memory

display. In the Number condition, we directly knew how many squares the participants thought there were. In the Location condition, by counting how many location responses they made, we got an indirect, but still informative measurement.

2.1.2 Result

I mainly focused on the number responses generated in the Number and Location conditions. We first kicked out number responses that were highly probable due to response errors (larger than 15 or equaled to 0, about 0.5% of the trials for both conditions). We then calculated the average number responses for different memory loads in each condition separately. The results are plotted in **Figure 2.3** and **Figure 2.4**. In **Figure 2.4**, the saturation level of a cell is shaded in proportion to the count for a given number response. From the two figures, it is clear that although the average number responses were roughly accurate, the standard deviations were large, suggesting that participants were uncertain about the number of presented objects in the memory displays. In accordance with previous studies, here, participants tended to underestimate than to overestimate when they generated a wrong number response (Izard & Dehaene, 2008; Ma & Flombaum, 2013).

To quantitatively analyze these data, I further calculated the enumeration error rate (the proportion of trials the participants made wrong number responses) for each condition (Figure 2.5). Repeated measure ANOVA was run for the Number and Location conditions respectively, with memory load as the independent variable. There was a main effect of memory load in both the Number ($F(7, 161) = 45.1$, $p<0.001$) and the Location condition ($F(7, 161) = 71.6$, $p<0.001$). The linear trend contrasts were significant for both conditions (Number: $F(1, 23) = 135.6$, $p<0.001$; Location: $F(1, 23) = 251.5$, $p<0.001$),

suggesting that participants became more and more uncertain about the number of objects as memory load increased.

16 one-sample t-tests were run to see whether these enumeration errors at different conditions were significant different from 0. With a Bonferroni correction (p=0.05/16), for the Number condition, enumeration errors were significantly higher than 0 starting from memory load of five, and for the Location condition, the starting point was two.

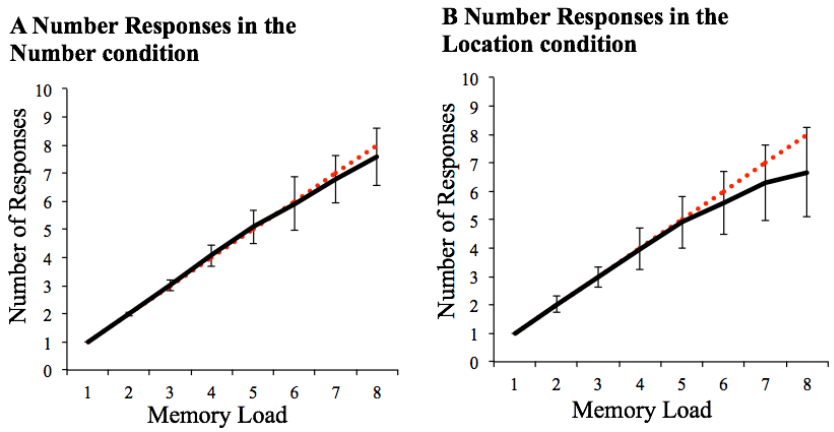


Figure 2.3. Average number responses as a function of memory load in A) Number condition and B) Location condition. Error bars show the standard deviation of number responses.

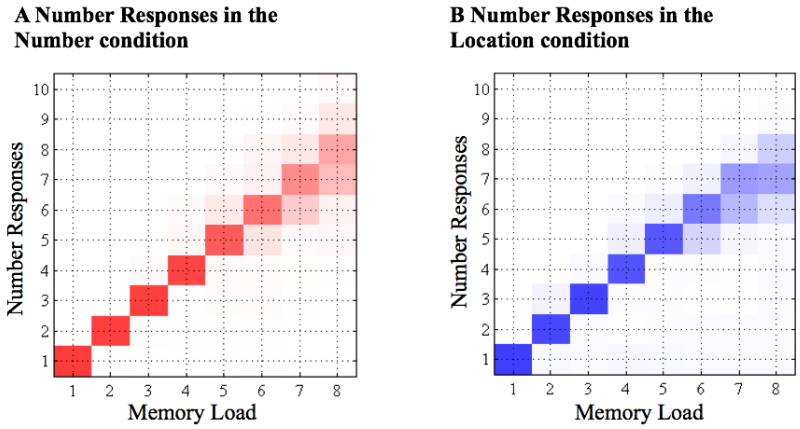


Figure 2.4 Distributions of the number responses made given a memory load. Distributions for the Number and Location conditions are shown in separate panels. The saturation levels of the cells are in proportion to the count of a given number response.

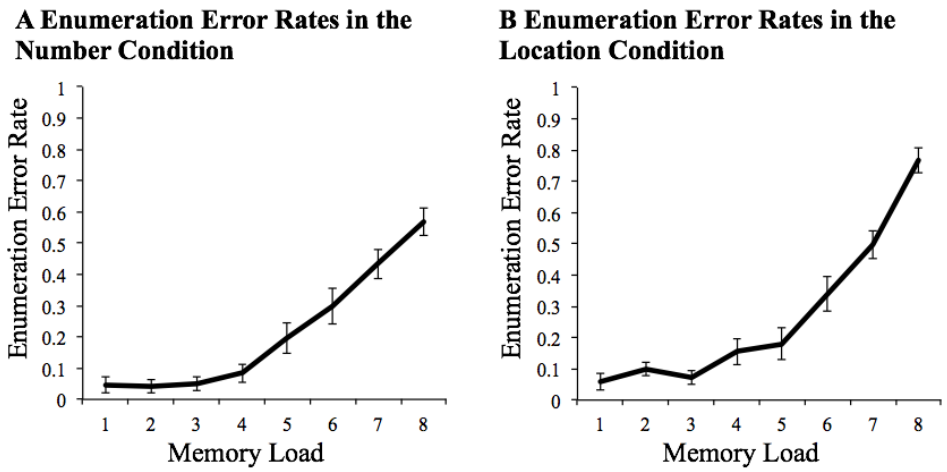


Figure 2.5 Enumeration error rates in the Number and Location condition as a function of memory load.

2.1.3 Discussion

In this experiment, I used two reporting methods to show that participants were uncertain about the number of to-be-remembered items in a standard memory display. In one condition, participants were directly asked to report the number of objects they saw. In another condition, participants were asked to click all locations where they saw objects. The number of responses they made was taken as an index of their representations of the number of objects in the display. In both conditions, participants tended to make more enumeration errors at higher memory loads. They sometimes overestimated, and sometimes underestimated. But in general, in accordance with previous studies on number perception, our participants tended to underestimate more than overestimate (Izard & Dehaene, 2008).

Our participants also made slightly more enumeration errors in the Location condition than the Number condition. In fact, with an MOT task, Ma and Flombaum (2013) also found the enumeration error rates in the clicking location conditions was

higher than in the reporting number condition. This difference could be caused by participants' aversion to guess or memory decay during the clicking process. Despite the difference between the two conditions, the critical finding is that the distributions were comparable. The results from the two conditions, taken together, could serve as evidence of people's uncertainty about the objects' presence in the memory display.

In sum, in this experiment, I combined a standard delayed estimation working memory task with two number report tasks. I showed that under a classic working memory task setting, people were uncertain about the number of objects presented in the display. The enumeration error rates increased to a very high level at higher memory load conditions. This reflected the possibility that object individuation is not always perfectly completed in typical memory displays.

2.2 Computational Experiment 2: A clustering model for correspondence assignment of samples in VWM

In this experiment, I continued to explore how the observed enumeration errors arose. I hypothesized that noisy correspondence computations might lead to some errors in signal-source assignment, and thus lead to inaccurate object individuations. Over- and under-estimation will happen when signals are not assigned to their real sources.

More specifically, I proposed a computational model that took uncertainty in stimuli representation and data assignment into account. By using a clustering algorithm, the model shared some key properties with human correspondence computation: both of them are based on noisy representation and are guided by spatial proximity. To foreshadow the result, the proposed computational model was able to approximate

people's performance with the same task. This model showed that the difficulty in signal-object assignment could be partially understood by a density-based clustering problem.

2.2.1 General model framework

Working memory representation doesn't start directly from objects. To simulate human correspondence process, the model first needs a way to represent the raw signals generated by the objects. In the current model, this process is implemented by generating random samples from each object. It is worth noting that we don't think there are little samples for each object in the human mind. This is just a good way to approximate the basis of noisy memory representation. Then, the model uses a clustering algorithm to assign noisy signals into different objects.

Step 1: Sample Generation

The model generated 50 random samples from noisy distributions centered at the true location of each item in the memory display (see **Figure 2.6** for an illustration of samples generated from one memory display). The location of the samples for a certain object i were drawn from the following two-dimensional normal distribution $N(\mu_i, [\sigma_i, 0, 0, \sigma_i])$, where μ_i was the x and y coordinate of the object, and σ_i was the standard deviation of the distribution. σ_i was dependent on the distance of the object to the current gaze location (we assumed the participant was fixating at the center of the display) and could be calculated by $\sigma_i = 0.2(1+0.42D_{i\text{-}fixation})$ (Carrasco & Frieder, 1997; Vul et al., 2009). Here, considering the additional noise induced by the memory interval, we chose 0.2 instead of 0.08 as the value of the standard deviation at fovea. The color of the samples for a certain object i were drawn from a von Mises distribution (circular normal distribution) that centered on the true color of the object (VM (Color$_i$, κ)). We used 14.89

as the value of the concentration parameter κ. This value was the average memory precision when only one item was asked to remember (Bae et al., 2015).

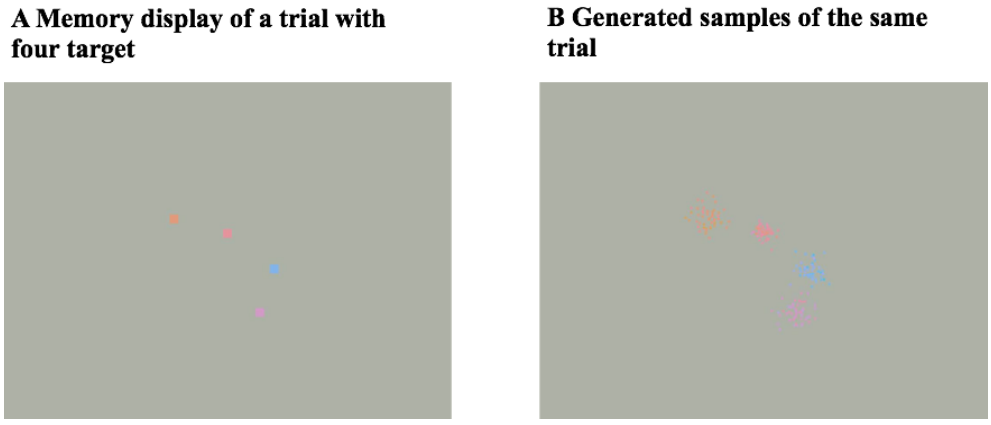


Figure 2.6 A sample memory display (A) and an illustration of model generated samples based on the memory display (B).

Step 2: Clustering

I sought to use a clustering algorithm to explain the difficulty in signal-object assignment. The model used a clustering algorithm to cluster all of the generated random samples and made a decision of how many objects were presented in the memory display. It applied a modified version of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN, Ester et al., 1996) algorithm to decide what and where were the most probable objects that can generate these samples. For my purposes, DBSCAN is an ideal algorithm most importantly because it did not require pre-defining the number of clusters (objects) in the display. Therefore, it could simulate the uncertainty both in the location of the objects, and in the number of objects. The DBSCAN algorithm is also very good at detecting areas that have high density and then group samples based on local density. Furthermore, it is generally good at detecting outliers, and thus it would be stable when some extreme samples are generated.

Two parameters, ε and MinPts need to be determined in a classical DBSCAN algorithm. ε-neighborhood of a sample is defined as the combination of all of the neighboring samples that are within ε distance of the current sample. A sample is defined as a core point if the number of its ε-neighborhood samples is at least MinPts. Sample p is called directly density-reachable to sample q when q is a core point and p is within ε distance to q. The clustering algorithm could be realized by the following iteration steps (See **Figure 2.7** for a simplified flow chart of the algorithm):

Step 2.1: Visit a random unvisited sample (P) generated in Step 1, if it is a non-core point, mark it as noise and restart Step 2.1. If it is a core point, start a new cluster and expand the cluster to all of the sample's ε-neighborhood samples. In other words, P and its ε-neighborhood are marked as members of the current cluster.

Step 2.2: Continue expanding the current cluster by including all ε-neighborhood samples of the current core point members. In other words, non-core points could be included in a cluster, but a cluster could not be expanded from a non-core points. Note that samples that are previously marked as noise could also be included as a cluster member in this step.

Step 2.3: When all core points of the current cluster have been expanded, repeating Step 2.1 to find new clusters.

The output of this algorithm contains: 1) The number of estimated clusters (objects); 2) The centroid of these clusters; 3) which samples belonged to which clusters, and which samples were treated as noise/outlier.

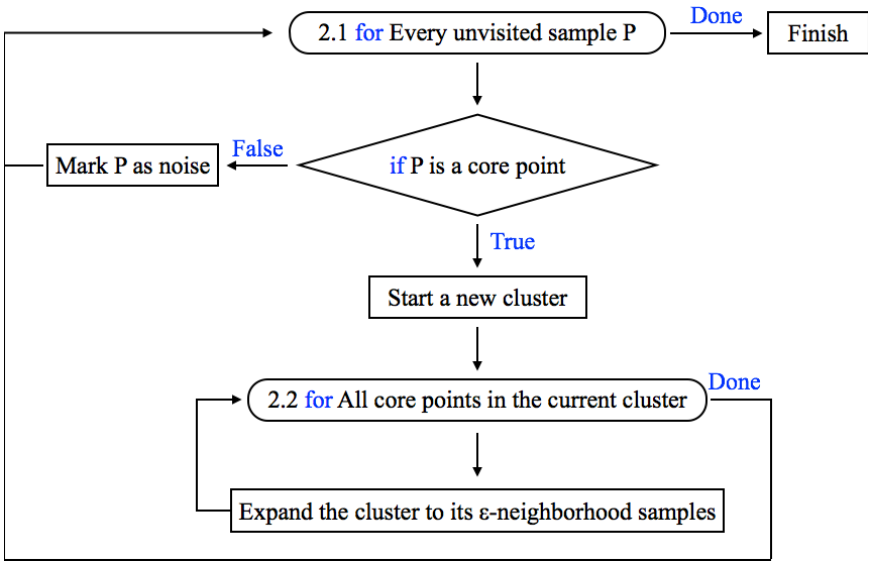


Figure 2.7 A flow chart of the DBSCAN algorithm

Parameter selection

It has been suggested that MinPts should be larger than 1 plus the number of dimensions D in the data set (Goss & Nitschke, 2014). Our data had two dimensions, x and y coordinates of a sample. Therefore, MinPts should be larger than 3. Since larger MinPts values would leave many samples marked as noise (Ester et al., 1996), we chose MinPts to be 4 in our model.

The value of ε determined the density of a cluster. The larger the ε, the higher the tendency that all samples would be clustered together (Ester et al., 1996). On the other hand, if ε had a too small value, many samples would be marked as noise. The ideal value of ε varied from dataset to dataset, with a suggested constraint that ε should take values between the minimum and maximum distance of samples to their nearest neighbors (here 0.01 and 1.01 degree, Goss & Nitschke, 2014). In the current model, we selected ε to be twice of the average standard deviation of the two-dimensional normal distribution we used to generate the samples. Our logic was that a point that didn't have enough

neighbors within two-standard-deviation distance isn't qualified as a core point to expand a cluster. Note that the standard deviations were dependent on the objects' distances to the fixation location. In order to use the same ε for all samples to be clustered, we first calculated the average distance to fixation location for all possible objects in our experiment, which was 3.72 degree. We then used this distance and computed the average standard deviation value, which was 1.02 degree. This value was just around the border of the suggested constraint of the ε value. Therefore, we used 1.02 as the ε value in our model.

The No-Color model and the Color model

I ran two versions of the clustering model, a No-Color version and a Color version. The No-Color model used the algorithm described above to cluster generated noisy samples into meaningful clusters, and treated the samples in one cluster as from the same objects. Color information was ignored in this model. In other words, a pink sample and a green sample could be clustered as from the same object, as long as they were spatially close to each other.

The Color model was the same as the No-Color model except for one additional constraint in the process of finding ε-neighborhood samples. In the Color model, in order to be counted as an ε-neighborhood sample of a specific sample P, a sample needed to satisfy both of the following criteria: 1) closer than ε to sample P, 2) be in the same color category as sample P.

The color categories of the samples were calculated based on the data reported in the category naming experiment of Bae et al. (2015). In that experiment, participants were asked to give color category names to all of the 180 colors used in the current

experiment, one each time. Participants could choose from six category names: “Pink”, “Orange”, “Yellow”, “Green”, “Blue” and “Purple”. The probabilities of different colors to be called as different categories were computed for further use.

For our purposes, we defined a color A to be in category I if on more than 30% of the time, A was named as I. We used 30% instead of larger values because for some boundary colors (e.g. a greenish blue), they could be given names of two categories with similar probability, but neither of them exceeded 50%. With a 30% criterion, one color could be counted in as much as three categories. Here, we defined “two samples in the same color category” as long as they shared any one of those categories (e.g. a greenish blue and a blue would be treated as being in the same color category).

2.2.2 Model testing method

We tested the Color and No-Color models with the same 320 memory displays we used for human participants in Section 3.1. For each trial, we first used the algorithm described in Step 1 to generate 50 random samples for each object in the original display. We did this 10 times for each trial to cover a reasonable range of randomness in the sample generation process. Then, we ran both the Color and the No-Color models independently with these samples, one time with each set of samples. We then used the number of clusters of each trial given by the models for further analysis.

2.2.3 Result

Similar to the analysis of human performance, we calculated the average number of responses for different memory loads in each condition separately. The results were plotted in Figure 2.8 and Figure 2.9. From the two figures, the No-Color model significantly underestimated the number of clusters. More importantly, both models

showed large variance in number estimation, suggesting that similar to human observers, the models were also uncertain about the number of presented objects in the memory display.

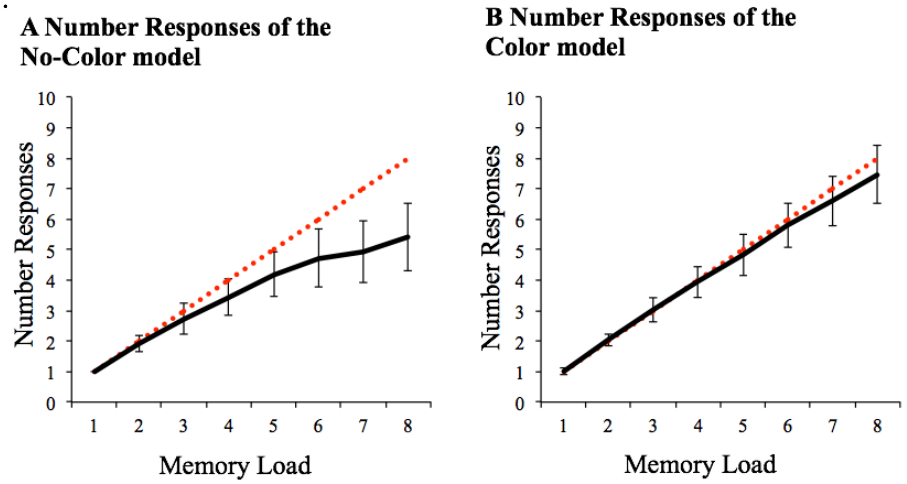


Figure 2.8 Average number responses as a function of memory load for the A) No-Color model, and B) Color model. Error bars show the standard deviation of number responses.

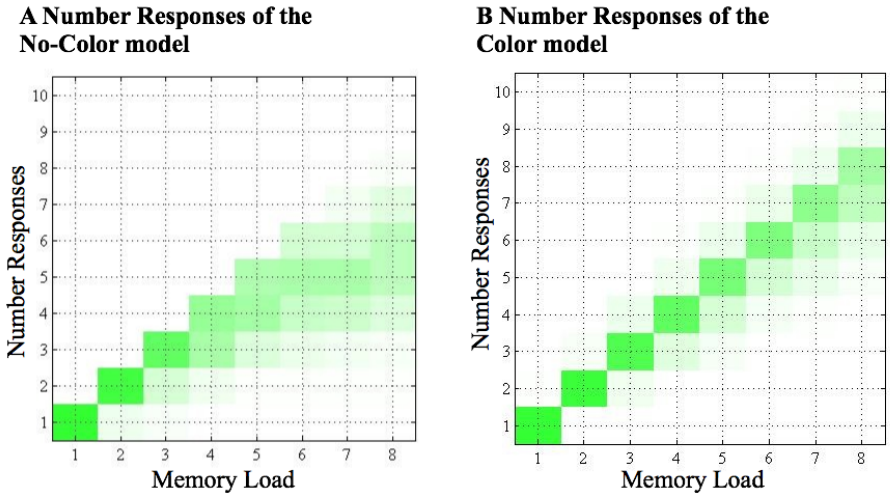


Figure 2.9 Distributions of the number responses mad by A) the No-Color model and B) the Color model, given memory loads. The saturation levels of the cells are in proportion to the count of a given number response.

We also calculated the enumeration error rate (the proportion of trials the participants made wrong number responses) for each condition (**Figure 2.10**). We treated each time of simulation as a participant (10 in total) and ran repeated measure ANOVAs for the No-Color and Color models respectively, with memory load as the independent variable. There was a main effect of memory load for both the No-Color ($F(7, 63) = 800.4$, $p<0.001$) and the Color model ($F(7, 63) =135.4$, $p<0.001$). The linear trend contrasts were significant for both models (No-Color: $F(1, 9) = 17271$, $p<0.001$; Color: $F(1, 9) = 690.5$, $p<0.001$), suggesting that similar to human participants, the model also became more and more uncertain about the number of objects as memory load increased.

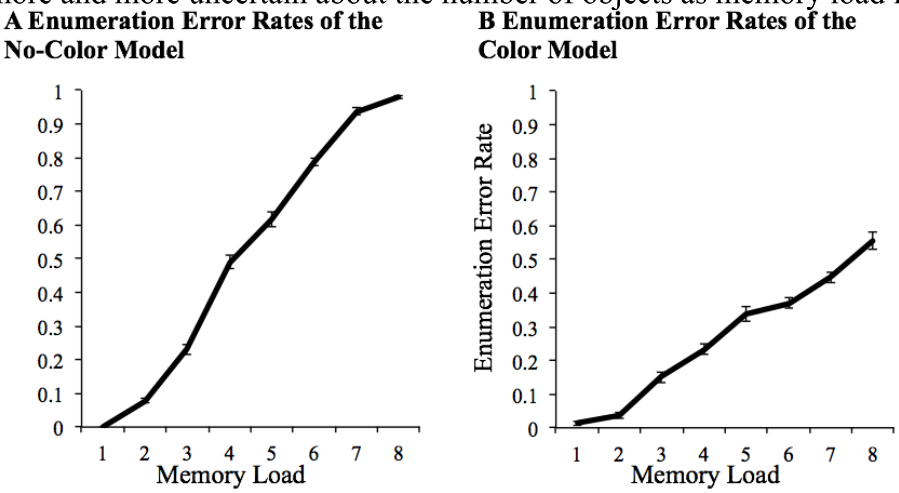


Figure 2.10 Enumeration error rates of A) the No-Color model and B) Color model as a function of memory load.

To further compare the similarity between human participants and our model, we ran a trial-by-trial partial correlation between the number responses generated by human and models, with target load as the controlling variable. This is to say, for all of the 320 trials, we first calculated the averaged number responses across all human participants, as well as across all model simulations. We then correlated these two sets of 320 values with target load as the controlling variable. This analysis allowed us to see whether the model

could capture the trial-by-trial variance (within each target load condition) we observed among human participants.

First, there was a significant positive correlation between the human number responses in the two testing condition, $r(317) = 0.186$, $p=0.001$. This result suggested that people were not just making random number responses. They were using the information in the display. There were inherent properties of each memory display that would lead to participants make more or less number responses.

We then correlated the two human measurements with the output of the two models respectively. For the No-Color model, there were significant positive partial correlations between the model number responses and the human number responses in both the Number condition ($r(237) = 0.192$, $p = 0.001$) and the Location condition ($r(237)=0.266$, $p<0.001$). For the Color model, there was a significant partial positive correlation between model and human responses in the Location condition ($r(237) = 0.144$, $p = 0.01$). In the Number condition, the partial correlation did not reach significance level ($r(237) = 0.088$, $p = 0.118$).

In sum, the models did show a similar behavior pattern to human observers. Especially in the Location condition, both the Color and No-Color model could capture the trial-by-trial variance in number responses, probably elicited by the properties of the memory display.

2.2.3 Discussion

In this computational experiment, I developed an algorithm that considered the uncertainty of both object properties and the presence of objects. Instead of representing each object independently, the model represented noisy samples generated from the

objects. A modified DBSCAN clustering algorithm was then used to identify the most reasonable sources of these noisy signals. Due to noise in spatial locations and color representations, samples from different objects could be mixed together, and samples from the same objects could be treated as from different objects.

This model showed similar number responses to human observers. First, the model showed increased performance variability with greater memory loads. With more objects in the memory display, both human observers and our model became more uncertain about the number of objects that could generate the received noisy signals.

Second, the No-Color model showed more enumeration errors and underestimation errors than human observers. This suggested that human observers used color information to segment the memory display.

Finally, the model's responses were significantly correlated with those of human observers. There was a small but significant correlation between the two conditions of human observers: human made similar number responses to the same memory display under two different reporting conditions. This consistency within human observers suggests that certain properties of the memory display led to either over- or under-estimation of the object number. This result itself further supports the idea that working memory tasks are not only about the uncertainty of the properties of objects, but also about how to form theobject representations from noisy signals. The next critical question was whether the model could capture this variance among different memory displays. In our computational experiment, the answer was yes. Both the No-Color and the Color model successfully captured the trial-by-trial variance in the Location condition. For the Number condition, the No-Color model did a better job than the Color model.

The current results suggested that human signal-source assignment process could be understood by a density-based clustering algorithm. However, the details and the specific ways to implement these algorithms might be different between human and our model. For example, the Color model failed to capture the trial-by-trial variance in the Number condition, suggesting that human observers might use the color information differently from the model. It is also possible that in addition to the correspondence computations, the enumeration task also involves other cognitive abilities such as the approximate number system (ANS, for reviews, Brannon, 2006; Feigenson et al., 2004). It has been proposed that ANS does not require accurate individuation of objects and is representing approximate cardinal values of sets objects (Feigenson et al., 2004). Since the Number condition is very similar to many tasks used to study the ANS, it's unsurprising if participants were partly relying on the ANS to generate their number responses. Future studies are needed to see how the ANS could be incorporated into the current model to better account for human behavior.

2.3 General discussion

Many previous visual working memory models and theories often assumed that all objects are perfectly individuated and represented in a VWM display. Participants are certain about the number of to-be-remembered targets (e.g. Zhang & Luck, 2008; van den Berg, et al., 2012). I designed the current study to understand how object representations are formed in memory, and how inaccurate these representations could be. I first used a behavioral experiment to show that people do not often have a perfect knowledge of the presence of all objects, which could be understood as the effect of inaccurate signal-

source assignment process. I then used a computational model to simulate possible processes used by human that can generate the observed behaviors. I found that a model that considered uncertainty in spatial representation, color representation, and a density-based clustering algorithm behaves very similar to human observers in terms of enumeration errors. These results support that signal-source assignment is a critical process in formation memory representations of objects, and could possibly further limit human memory ability.

These results are also consistent with previous findings that people sometimes make responses to nontarget object in delayed estimation tasks (Bays, Catalao, & Husain, 2009). In fact, non-target based response should arise naturally if signals are not correctly assigned to their sources. The non-target based responses could also not be only based on nontarget, but be based on a mixture of signals from multiple source objects. Further studies are needed to study more precisely about the kind of color response errors people make in a VWM task.

Another important finding in the current study was that people made consistent number estimations under different reporting conditions. Certain properties of the memory display, e.g. the spatial and color properties, therefore must lead human observers to consistently over- or under- estimate the number of objects in the display. The fact that our model could show similar trial-by-trial responses to human observers suggested that human were using similar algorithms when they are clustering the memory display.

There are a lot of similarities between our study and many studies on perceptual grouping. Perceptual grouping happens when observers organize visual inputs into

distinct objects or clusters (Feldman, Singh, & Froyen, in press). There are two levels that our study could be similar to a perceptual grouping study. First, participants could group individuated objects into different clusters. Second, for our model, the clustering process could also be treated as a perceptual grouping process, where the model groups similar samples together.

Orhan and Jacobs (2013) proposed a model that involved grouping at the object level. In their study, they developed a nonparametric Bayesian mixture model to cluster distinct objects in a memory displays to different groups. They found that this model could produce similar memory bias to human observers. However, one implicit assumption of their model was that the clustering and organization was based on perfectly segmented objects. In other words, though they considered the possibility of clustering objects, the clustering was still based on the raw objects stimuli. The model ignored the critical step of image segmentation and data assignment. This step seems to be trivial, but is actually very important. For example, since their model was based on object, their model would never over-estimate the number of objects in the display, which happened a lot in our data. Therefore, simply considering clustering, but ignoring the image segmentation step, is not enough for a working memory model.

At the level of the noisy samples, our model shared some similarities with many other perceptual grouping models. For example, Im, Zhong, and Halberda (in press) used a modified k-means clustering algorithm and found that people might have a default 4° visual angle window to group similar dots together. Moreover, as more dots were clustered together, human observers made more underestimation errors on the number of presented dots. This study provided robust evidence that perceptual grouping is based on

proximity. This model is more suited to explain how people are grouping randomly generated similar dots into clusters. In our case, there is inherent relationship among different samples, and color added another dimension for the clustering process. Therefore, it's hard simply apply their model to our data. However, they provided converging evidence that the properties of different displays could lead participants to consistently over- or under- estimate number of objects on the display. These results suggest that considering the algorithm of image segmentation process is very important in analyzing human behaviors.

Feldman, Singh, and Froyen (2014) proposed a Bayesian mixture model to estimate the number of groups in a display of multiple dots. Instead of a density-based nature, they used Bayesian estimation to estimate the sources (hypothesized objects) that have the maximum posterior probability to generate the noisy dots. Similar to the DBSCAN algorithm, this approach did not set a hard constrain on the number of clusters. By selecting appropriate priors, the model will automatically favor outputs with fewer sources. We think both our density-based algorithm and their Bayesian algorithm could serve as a good approximation to human clustering process. Since their paper did not provide a direct comparison between model and human performance, future work is needed to compare model performance and see which algorithm(s) could better explain human behaviors. More importantly, we are not arguing that people are using a density-based algorithm to assign signals to their sources. What we want to show is that a model that do data assignment in some way could behave very similar to human, and thus data assignment and image segmentation should be treated seriously in any formal model of VWM.

In sum, in this study I showed that visual working memory amounts to more than estimating the properties of objects. It also involves inferences about the presence of objects and therefore signal assignments. This added level of uncertainty places independent constraints on the capabilities of visual working memory. Future work should focus on comparing different clustering algorithms, as well as apply the model to simulate human memory responses.

Chapter 3: The role of transient signal detection and object-based correspondence computation in motion perception

(Some parts of this chapter were previously reported in Ma, McCloskey, & Flombaum (2015), A deficit perceiving slow motion after brain damage and a parallel deficit induced by crowding)

Perceiving motion inherently involves correspondence computation. It's very important for the observer to assign signals received across time to the same source to perceive a coherent motion trajectory. Accurate correspondence computation is even more important for the participants to correctly detect the direction of a moving stimulus. If the representations of the signals are too noisy, or signals are assigned to wrong sources, participants can not make accurate judgment of motion direction.

There are at least two correspondence algorithms can be implemented to make motion perception possible. The two systems are often referred to the higher level and lower level systems respectively (Battelli et al., 2001; Lu & Sperling, 1996, 2001). There are open debates and questions concerning the exact inputs, algorithms, and neural structures that support each system. Broadly, a lower level system automatically and preattentively take transient signal changes as its only inputs (Adelson & Bergen, 1985; Hock, Gilroy, & Harnett, 2002; Reichardt, 1961; van Santen & Sperling, 1984). Some neurons in the primary visual cortex show these signatures, as do neurons in the middle temporal/medial superior temporal (MT/MST) areas (Britten & Heuer, 1999; Tootell et al., 1995). Correspondence relationship is automatically computed as long as two temporally separated signals can activate a neuron in this system. A higher level system generally relies on focal and object-based attention and implements explicit algorithms to make correspondence analysis between object positions (Burr & Thompson, 2011;

Cavanagh, 1992; Dawson, 1991; Petersik, 1995; Ramachandran & Anstis, 1986; Seiffert & Cavanagh, 1999; Ullman, 1984). The brain system supporting these computations probably involves areas that are higher in the hierarchy, including the inferior parietal lobe (Battelli, Pascual-Leone, & Cavanagh, 2007). For purposes of clarity and economy, I will refer to the preattentive detection of transient luminance signals as *lower level* and the representation of objects with engagement of selective attention as *higher level*.

In this chapter, I will use motion perception as a special case to study how these different correspondence computation algorithms are used in visual cognition. More specifically, since both the lower and higher level systems can generate coherent motion perception, it's important to see how they are interacting with each other during motion perception.

I will study the relative contribution of the two correspondence algorithms in two different types of motion stimuli: smooth (modal) motion and apparent (amodal) motion. We perceive the motion of modal stimuli: objects that remain continually in view. We also complete trajectories despite amodal moments—for instance, when an object becomes occluded (Burke, 1952) or when objects rapidly change position noncontiguously, even across relatively long distances (Braddick, 1974; Petersik, 1989).

It remains unclear exactly how the two motion systems are involved in perceiving the direction of smooth and apparent motion. It seems likely that higher-level correspondence systems are necessary in at least some amodal conditions—for example, when noncontiguous position shifts are large and infrequent. It's also intuitive to think that the lower level correspondence system is involved in the perception of smooth motion, since "a rigidly moving object is a drifting modulation of luminance" (Lu &

Sperling, 1996, p. 44). However, does this mean the higher-level system is not necessary for smooth motion perception, and the lower-level system is not needed for apparent motion perception?

Under normal conditions, both the higher- and lower-level systems are often generating the same output perception. This makes it very hard to study the relative contribution of the two systems. In this chapter, I will show that under some special cases, one of the two systems will not have enough signals to generate reliable outputs, or they will generate opposite outputs. These cases provide us unique opportunities to study the relative contributions of different correspondence computations in motion perception.

I will first investigate whether a higher-level correspondence computation is required for observers to perceive the direction of slow but smooth motion stimuli. This research question is motivated by a brain-damaged patient, JKI, who had symptoms consistent with an impaired object localization system. I will show that his deficit in localizing objects may further cause a failure in the correspondence computation and thus lead to difficulty in perceiving motion direction of slowly moving objects. I will then report the result with healthy participants to confirm the hypothesis that the perception of slow, smooth motion depends on higher level correspondence computation of object positions across time.

The second goal of the research is to study how the lower level transient detection system and the higher level correspondence system are involved in the perception of amodal motion stimuli. Under certain viewing conditions, the two systems will generate contradicting explanations of the same apparent motion display. I will take advantage of this property to study the relative contribution of the two systems in perceiving apparent

motion.

3.1 Experiment 3: Patient JKI's spatial-localization deficit

3.1.1 Method

Patient history

JKI is a 51-year-old right-handed male with a college education who suffered multiple strokes in 2003. MRI in 2011 revealed bilateral damage (see **Figure 3.1**). The damage was more extensive on the right, affecting lateral and inferior surfaces of the temporal lobe, much of the parietal lobe, and the lateral occipital lobe. Damage in the left hemisphere was largely restricted to the posterior parietal lobe and superior occipital lobe. Primary visual cortex was spared in both hemispheres. The right-hemisphere damage includes middle temporal/medial superior temporal (MT/MST) areas, which have been associated with motion signal detection (Tootell et al., 1995).

As a consequence of the right-hemisphere damage, JKI suffers from partial paralysis of the left arm and leg. He also has visual field defects, showing impaired detection of stimuli presented in the lower left visual field, and in the medial portion of the lower right visual field. JKI also reports frequent diplopia (double vision) with binocular but not monocular viewing.

Neuropsychological testing revealed intact language and memory, but significant visuo-spatial deficits characteristic of patients with bilateral parietal damage (e.g. Robertson, Treisman, Friedman-Hill, & Grabowecky, 1997; di Pellegrino & de Renzi, 1995; Humphreys & Riddoch, 1993). In particular, JKI is severely impaired in copying simple pictures or designs, producing fragmented and inaccurate copies; he is impaired in reaching for visual targets in the upper left visual field, despite being able to detect the

targets; and he shows extinction/simultanagnosia for upper left visual field targets, often failing to report a target if another target is presented further to the right. Some of these deficits will be described in greater detail below.

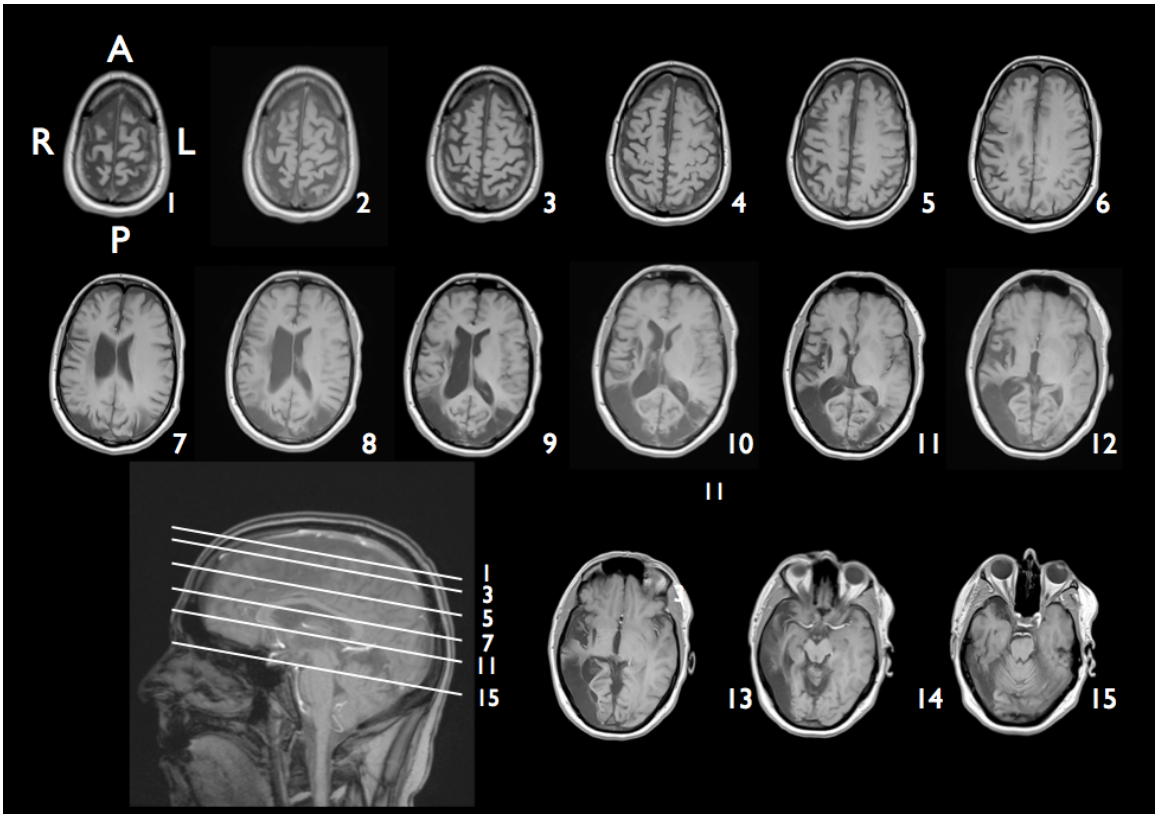


Figure 3.1 T1 weighted magnetic resonance (MR) images of JKI's lesion. Top panel displays a series of axial images. Lower panel displays the location of these slices on a sagittal image.

Apparatus and test setting

All testing with JKI took place in a dim room. Stimuli were generated with MATLAB and the Psychophysics toolbox (Brainard, 1997; Pelli, 1997) and were presented on a MacBook Pro laptop with a refresh rate of 60 Hz. The viewing distance was fixed to 55 cm so that the whole display subtended 29.6×18.5° of visual angle. A chin-rest helped JKI maintain a stable posture and fixation.

Fixation during testing was monitored informally, and corroborated by a small number of eye-tracker sessions. Eye-tracker calibration and testing were strenuous for JKI, which limited the ability to collect sufficient data. But performance appeared qualitatively similar with fixation maintained during those sessions.

Because of JKI's lower visual field defects, all stimuli in the experiments we report were presented in the upper half of the visual field. Also, to avoid double vision during testing, JKI's right eye was patched throughout all of the experiments.

Procedure

I investigated JKI's ability to detect and localize objects in the upper visual fields. On each trial either one or two objects were presented for 750ms. JKI was asked to report the color and shape of each object seen during a trial.

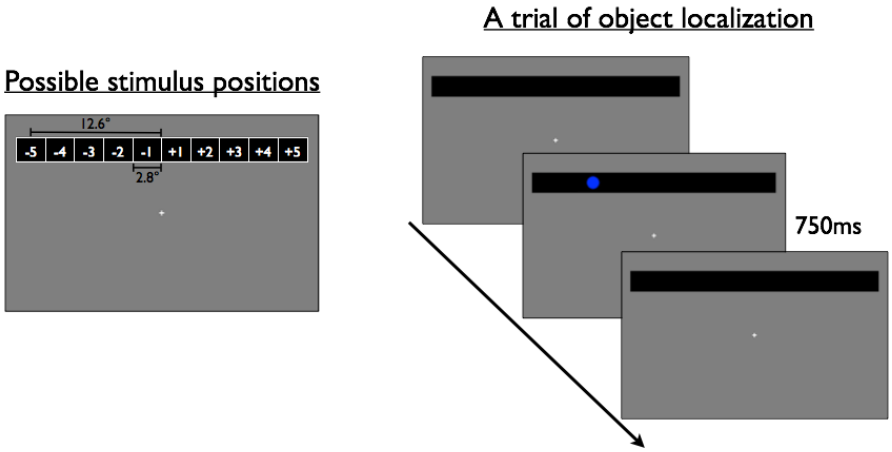


Figure 3.2 Procedure employed for investigating JKI's object-representation deficits. Each trial included one or two colored shapes that appeared in ten possible positions. JKI's task was to report the shape, color, and location of any stimulus objects he observed in a trial.

Additionally, he was asked to report the location of each object. The objects were constrained to a black strip 6° above fixation. The strip (27.7° × 2.31°) was further divided into ten equally spaced horizontal positions (2.77° apart from one another,

measured center to center, **Figure 3.2**). Objects always appeared in one of these ten locations, and in two-object trials, the objects always appeared in two different positions (and with different shapes and colors). JKI was familiarized with the numbering of these ten locations from -5 to +5, as pictured in the figure, and his understanding of this numbering system was confirmed via testing with free viewing (i.e. without fixation) as well as with a tactile apparatus. Position numbers were not displayed during the experiments.

3.1.2 Results

Figure 3.3 graphs the likelihood that JKI correctly reported the color and shape of an object as a function of its position on the screen in the single-object trials. JKI was able to accurately report the shape and color of an object presented in the upper right visual field. In the upper left visual field, his performance was also relatively good (mean accuracy = 95%), though he occasionally missed some objects presented at positions -5 and -3. These results suggested that JKI has little or no deficit *detecting* a single, briefly presented object in either of the upper visual fields. (With two objects presented simultaneously, performance in the left visual field showed evidence of extinction or simultanagnosia. In particular, JKI frequently failed to report the leftmost stimulus when that stimulus appeared in the left visual field. The implications of the two-object results are outside the scope of this report.)

I also analyzed JKI's position reports in single object trials. **Figure 3.4** shows the average location at which JKI reported an object as a function of its actual location. He demonstrated an inability to accurately report object locations in the left visual field, even when he could accurately report the colors and shapes of the relevant objects. His deficit

was systematic, with objects generally reported farther towards the vertical meridian (i.e., to the right) than they actually appeared. In the upper right visual field, in contrast, position reports were generally accurate.

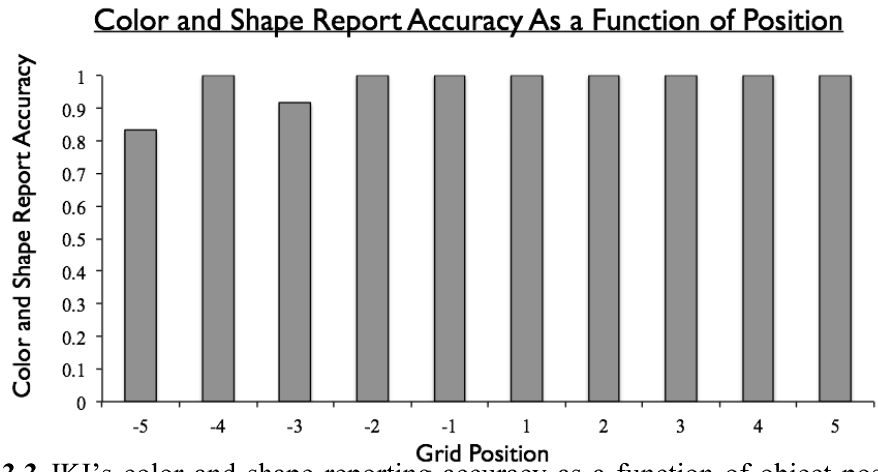


Figure 3.3 JKI's color and shape reporting accuracy as a function of object position in single-object trials. Objects were randomly presented in one of ten positions. Accurate responses involved reporting both the shape and color of an object that appeared. Here results are plotted as a function of an object's actual position, though this does not imply that JKI reported the relevant object as being in that position.

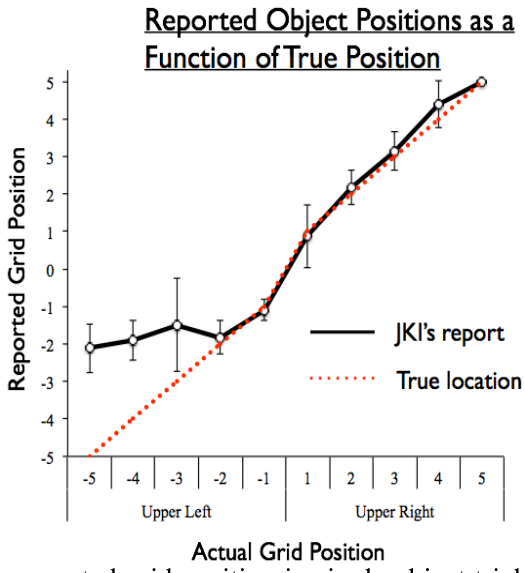


Figure 3.4 JKI's average reported grid position in single object trials as a function of an object's true position. Error bars show standard deviations (nine trials per position).

3.1.3 Discussion

For current purposes, the main implications of these initial evaluations are that JKI appears to have a deficit of object-position representation in the upper left visual field. In the upper right visual field, however, he appears largely unimpaired. Since the correspondence computation in the higher-level system involves comparison among object locations across time, JKI should have problem in using the higher-level system in the upper left visual field. This led me to investigate JKI's motion perception abilities, reasoning that in the upper left visual field he should show impairment in any motion perception task that relies on an intact higher-level system for representing object position. In contrast, he might possess preserved abilities with motion perception that arises more directly from lower-level signals.

3.2 Experiment 4: JKI's smooth motion perception

3.2.1 Method

To test JKI's ability to perceive the direction of continuously moving stimuli, I conducted an experiment in which, on each trial, a white disc moved a fixed distance of 1.39°. The disc's speed varied by trial; I utilized six different speeds (1.95°/s, 0.97°/s, 0.49°/s, .32 °/s, .24 °/s and 0.19°/s) to cover a wide range, including two relatively fast speeds (around 1°/s or greater) and four relatively slow speeds (less than .5°/s).

The disc subtended 0.9° in diameter and had a luminance of 198 lx. It could move either to the left or right. We used three anchor positions (4.16°, 5.54°, and 6.93° away from central fixation cross, around the ±2, ±3, and ±4 grid positions in the localization experiment) as the starting and ending positions of motion in each visual field. On each

trial motion was restricted entirely to either the left or right visual field; stimulus objects never crossed the vertical meridian. For each trial, two adjacent positions (e.g. 4.16° and 5.54° to the left of fixation) were selected as the starting and ending locations. Throughout its horizontal motion the disc's vertical position was fixed at 6° above fixation. The background was black, producing strong contrast with the disc.

An experimenter initiated each trial after JKI indicated that he was fixated and ready. JKI's task was to report verbally whether the disc had moved to the left or to the right. Testing was divided into several blocks across several weeks. In total, JKI completed 24 trials with each speed in each visual field.

3.2.2 Results

JKI's directional judgment accuracy is shown in **Figure 3.5**. In the right visual field, his performance was nearly perfect (98.6%) and there was no significant effect of motion speed (χ^2 (5) = 7.15, p =0.21, Cramér's V= 0.22). In the left visual field, he showed a speed-dependent impairment, with worse performance under the slow speed conditions. A chi-square analysis showed a significant main effect of speed (χ^2 (5) = 18.16, p = 0.003, Cramér's V= 0.36). We then averaged his performance with the two faster speeds and compared with average performance with the four slower speeds. There was a significant difference between the two fast and the four slow speeds (χ^2 (1) = 16.90, p < 0.001, Cramér's V= 0.34).

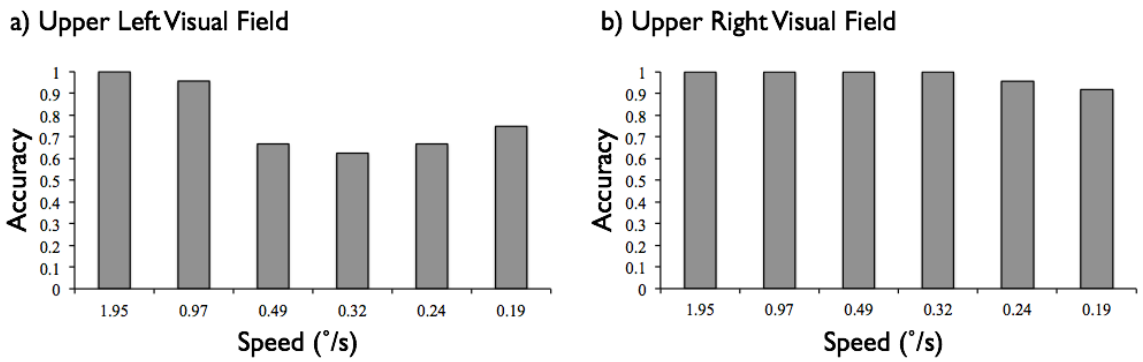


Figure 3.5 JKI's directional judgment accuracy as a function of speed in the upper left visual field (a) and the upper right visual field (b).

3.2.3 Discussion

In the left visual field, JKI's ability to judge the motion direction of a slowly but continuously moving local target was considerably impaired compared to his ability with faster moving targets, and with targets at any of the tested speeds in the right visual field.

Note that because the motion distance was fixed across all speed conditions, slower motion produced targets displayed for *longer* durations. Hence, JKI's impaired perception of slow motion occurred despite the longer durations over which to assess motion direction.

The motion perception results, in conjunction with JKI's deficit in position representation, are consistent with the hypotheses that JKI generally utilizes higher-level correspondence mechanisms to evaluate slow motion. For slow motion, the luminance change in the spatial-temporal domain is much less rapid than that for fast motion, and thus transient signals alone are not strong enough to reliably support low-level motion perception.

Of course, I cannot entirely exclude the possibility that JKI simply has unrelated position and motion deficits. Especially given the presence of damage in the right middle temporal/medial superior temporal areas, it is possible that JKI suffers from a motion-

signal detection impairment, which differentially interacts with motion speed. We are unaware, though, of any reports of hemifield-restricted low-level motion deficits induced by middle temporal/medial superior temporal brain damage; recall that JKI's performance in the right visual field appeared spared and did not interact with motion speed. Moreover, the intriguing possibility suggested to us by JKI's pattern of deficits is that higher-level mechanisms may usually play a compensatory role in slow, smooth motion perception, whenever low-level signals are impotent, whether because of extenuating factors such as brain damage, or because they are simply too weak.

I therefore sought to further explore the hypothesis that perception of slow continuous motion relies upon higher-level correspondence mechanisms, and that JKI's associated deficits reveal a dependency that applies in neurologically intact observers as well. Normal observers may not distinguish intuitively between faster and slower smooth motion because both are usually detectable. Below the surface, however, faster and slower motion may rely differently on lower- and higher-level mechanisms. Faster motion may rely on luminance changes (and may also benefit from redundancy with higher level mechanisms), while slower motion perception may rely more heavily on corresponding object positions across time, owing to gradually attenuated lower-level transient signals as speed decreases. I will investigate this hypothesis in the next experiment.

3.3 Experiment 5: A motion-perception deficit induced in healthy participants via crowding

JKI's deficits suggested the possibility that in general —that is, in healthy, visually normal observers— the perception of slow continuous motion relies heavily on

higher-level correspondence systems, perhaps because any transient luminance changes are separated by durations much larger than the typical settings of low-level detectors. A way to test this possibility is to impair high-level abilities in healthy observers, and then to determine whether a smooth, slow motion impairment emerges as a consequence.

Towards this end, I employed spatial crowding. For objects presented in the visual periphery, nearby flanker objects impair the ability to discriminate and localize individual objects, a phenomenon known as crowding (Intrilligator and Cavanagh, 2001; Pelli & Tillman, 2008; Whitney & Levi, 2011). The exact nature and causes of crowding-induced impairments remain debated (e.g. Levi et al., 2002; Pelli et al., 2004; Wilkinson et al., 1997). Some theorists have suggested, for example, that crowding impairs the ability to allocate object-based attention to individual objects, in turn impairing any mechanisms that rely on object-based attention, including individuation (e.g. Intriligator & Cavanagh, 2001; He, Cavanagh, & Intriligator, 1996). For current purposes, the important consequence of crowding is an impaired ability to localize individual objects accurately. And thus, under crowding, it should be harder for participants to accurately individuate object locations and thus make judgment of motion direction.

Via crowding, I sought to devise a motion judgment experiment as similar as possible to the one used with JKI. A white circle was shown to participants in the periphery, at one of four possible speeds, and under conditions with and without crowding inducers. The task was to report the direction of the moving target. I reasoned that if crowding makes it difficult to extract high-level information about object position and slow motion produces low-level signals that are too weak, then crowding should

impair motion direction judgments for slow stimuli (compared to uncrowded stimuli, where high-level information can be extracted).

3.3.1 Method

Participants

22 undergraduate students at Johns Hopkins University participated in this experiment for course credit. All had normal or corrected-to-normal visual acuity.

Stimuli and procedure

Stimuli were presented on a Macintosh iMac computer with a refresh rate of 60 Hz. The viewing distance was approximately 60 cm so that the whole display subtended 39.4×24.8° of visual angle.

The moving stimulus was a white disc that always moved a distance of 0.64° in a trial, at a speed of 1.53, 0.64, 0.14 or 0.08°/s. It was always presented in the peripheral field of one of the four possible quadrants (i.e. upper left, upper right, lower left and lower right), starting its motion 9° (diagonally) from the central fixation cross (see **Figure 3.6;** A demonstration of these conditions, and all those reported can be viewed online at http://www.jhuvisualthinkinglab.com/maetal-motiondeficit).

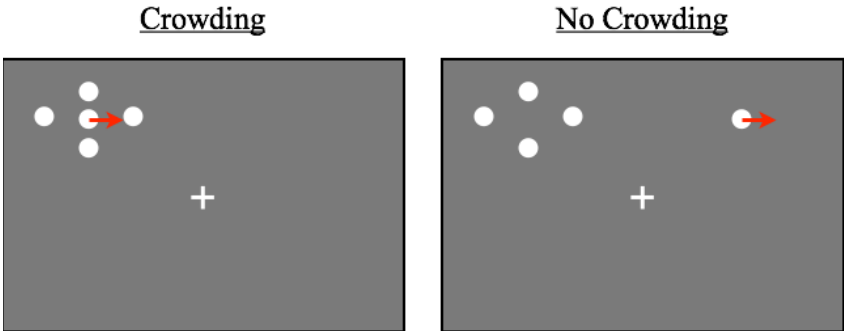


Figure 3.6 Schematic depiction of stimuli under conditions of Crowding and No-Crowding. The moving disc could appear in any of the four quadrants. Flanker objects were present in each trial, in the Crowding condition, in the same quadrant as (and surrounding) the moving object, and in the No-Crowding condition, in one of the remaining three quadrants. All of the flanker objects remained static during a trial. A demonstration of these conditions and all those reported can be viewed online at http://www.jhuvisualthinkinglab.com/maetal-motiondeficit).

In the crowding condition, four static flanker discs (identical with the moving one) were presented surrounding the moving target disc, as shown in **Figure 3.6**, at positions 2.39° to the right and left of the moving disc's starting position, and also 1° directly above and below. In the no-crowding condition, the moving disc was presented alone in one of the four possible quadrants. The four static flanker discs were presented in one of the other three quadrants.

Participants were instructed to fixate a central cross. Fixation was not monitored, but a secondary task was used to encourage fixation. At a random point during each trial, a single digit (between 1 and 9) replaced the fixation cross for 167ms. At the end of each trial, participants first pressed the 'left' or 'right' arrow key to indicate the motion direction they observed in that trial, and second, they entered the digit that had appeared at fixation.

Each participant completed 48 trials for each combination of crowding condition and speed, a total of 384 trials per participant.

3.3.2 Results

All trials with an inaccurate digit report were excluded from further analysis, which amounted to 8.7% of trials.

Motion judgment accuracy is presented in **Figure 3.7.** There was a significant main effect of speed condition ($F (3, 63) = 13.61$, $p < 0.001$, $\eta_p^2 = 0.39$), and a significant main effect of crowding condition ($F (1, 21) = 14.47$, $p = 0.001$, $\eta_p^2 = 0.41$). Critically, there was a significant interaction between speed and crowding ($F (3, 63) = 6.70$, $p < 0.001$, $\eta_p^2 = 0.24$).

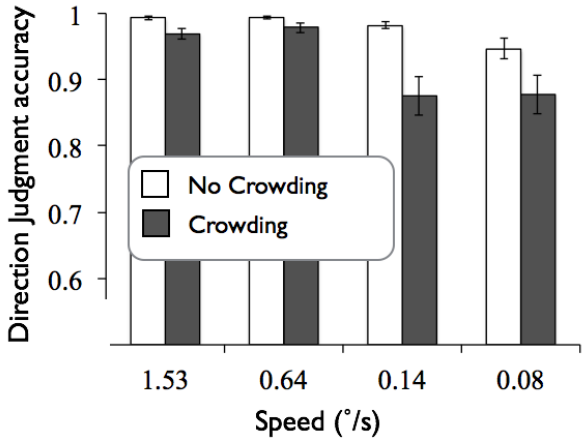


Figure 3.7 Directional motion judgment accuracy as a function of object speed and crowding condition, healthy participants.

To further scrutinize the effect of crowding, I looked at the simple main effect of crowding by speed. Crowding had a significant effect on performance at the two slower speeds (with Scheffé correction, at 0.08°/s, $F (1, 21) = 11.8$, $p=0.002$; at 0.14°/s, $F(1, 21) = 28.5$, $p < 0.001$), but did not significantly affect performance with the faster speeds (with Scheffé correction, at 0.64°/s, $F (1, 21) = 0.57$, $p=0.46$; at 1.53°/s, $F(1, 21) = 1.38$, $p = 0.25$). Participants performed significantly worse with crowded stimuli, but only with slower speeds.

3.3.3 Discussion

Judging the motion direction of a slowly moving peripheral target proved more difficult for healthy, visually normal observers when flankers crowded the slow moving target. Crowding is known to impair higher-level object individuation and localization abilities. Accordingly, these results are consistent with the hypothesis that the perception of slow, smooth motion requires higher-level object-based correspondence mechanisms, while faster motion can rely entirely on lower-level transients detection system that is available despite crowding, and that assigns correspondence relationships more implicitly.

This experiment suggested that when perceiving smooth motion stimuli, the two motion systems are to some degree complimentary to each other. The lower-level system efficiently detects quick luminance transients, while higher-level correspondence mechanisms can compensate by relying on representation of token position, and perhaps by exploiting object-based attention to improve spatial resolution. In the next experiment, I will further explore the interaction between the two systems when they are involved in perception of amodal apparent motion stimuli.

3.4 Experiment 6: Differentiating the contributions of lower-level motion system and higher-level correspondence system to perceptions of the Ternus display

The Ternus display (**Figure 1.3**) typically consists of two sequentially presented frames of objects. Both frames contain three equally distributed objects, but the position of the leftmost object in the second frame is the same as the middle object in the first frame. Under different conditions, the Ternus display is known to elicit two different perceptions: element motion (one object moving with two stationary middle objects) or

group motion (all three objects moving together, Petersik & Rice, 2006). Shorter interstimulus intervals (ISI) between the two frames often induce more element motion perception, while longer ISIs often induce more group motion perception.

Ternus display is useful for investigating correspondence computation because multiple correspondence interpretations are available in one display. It has been shown that the two perceptions are mutually exclusive such that participants do not report seeing both element motion and group motion at the same time. Adaptation effect also exists (Petersik & Pantle, 1979), such that after seeing one perception (e.g. element motion) a long time, participants will have higher probability to see the other perception (e.g. group motion). These results suggest that perhaps the two perceptions also involve the participation of different system for correspondence analysis.

Many theories have been proposed to explain how the two perceptions are generated. The most famous theory is the 'short-range' versus 'long-range' processes theory. It has been suggested that element motion is the output of the 'short-range' process (similar to the lower-level transient signal detectors) signaling zero movement for the middle elements, and the 'long-range' process signaling the correspondence between the outer stimuli across the two systems (Braddick & Adlard, 1978). In other words, to perceive an element motion in a typical Ternus display, one needs to first use the lower-level motion system to infer there is no motion energy for the middle elements, and then the higher-level correspondence system can only match the outer elements in the two frames together.

This theory could explain many typical phenomena observed in the Ternus effect. For example, with longer ISIs, the zero motion energy of the middle elements are also

weakened, and thus group motion is reported more. However, it does not explain why more peripheral presented Ternus display could induce more group motion (Breitmeyer & Ritter, 1986), and also why shorter presentation time of each Ternus frame could induce a third type of perception: simultaneity (a perception that all possible element positions are occupied at the same time, Dawson & Wright, 1994). Therefore, another theory based on the persistency of elements has been proposed. In this theory, it's the persistence, rather than the zero motion energy, that is determining the stationarity of the middle elements. The larger the persistence of the middle elements, the higher probability they will be integrated across time, and thus the higher probability element motion will be observed. Dawson and Wright (1994) further developed a computational model that can simulate human perception under different Ternus conditions. However, this persistence theory only focuses on the higher-level correspondence between the persisting signals, and completely ignores the possible involvement of the lower-level transient detection system.

Other explanations based on Gestalt grouping theories have proposed that if one can successfully group elements in one frame, group motion is favored. If one can group elements across frames, element motion is favored (He and Ooi, 1999; Kramer & Yantis, 1997). The problem with these theories is that they do not specify how grouping is done in simple Ternus displays where no specific grouping cues are provided.

Despite the intuitive nature of many theories of the Ternus effect, they all fail to explain some observed human behaviors (Petersik, 2006). The goal of the current study is not to provide a comprehensive explanation of the phenomena. On the contrary, I want to understand the effect under the framework of correspondence computation. I will explore

the contribution of different correspondence computation algorithms to the two perceptions of the Ternus display.

To study the relative contribution of different algorithms, it is very important to make sure different algorithms are generating different correspondence relationships of the same display. This is because if different mechanisms are predicting the same behavior, it will be really hard to know the relative contributions of each system. Taking a close look at the stimuli, this condition is satisfied only when the Ternus display is presented in the peripheral visual field.

When the Ternus display is presented around the fovea area, with a 0 ms ISI, the output of multiple correspondence analysis is element motion perception. First, the higher-level position comparison system can precisely localize the middle elements and thus correspond position 2 and position 3 to themselves using the nearest neighbor rule. Second, the lower-level transient detection system can easily compute there is zero motion energy at position 2 and position 3. Finally, the lower-level transient detection system could also detect that there is motion energy from position 1 to position 4. Therefore, all three analyses would make a consistent output that the outer element is moving back and forth.

When the ISI is increased for Ternus display presented at central vision, different correspondence systems will then start to favor group motion. First, the localization precision at position 2 and 3 will decrease due to the requirement of memory, or a lack of visual persistence as suggested by some researchers (Breitmeyer & Ritter, 1986). Therefore, the higher-level system will have a harder time to apply the nearest neighbor rule, and the relative velocity rule will be applied instead. Second, motion energy will no

longer be zero at position 2 and 3, while the transient signal change from position 1 to position 4 will be much weaker. Therefore, the output of both the higher and lower level systems will show a decrease in element motion perception and an increase in group motion perception.

The two systems will have different correspondence outputs when the 0ms ISI Ternus display is presented in peripheral vision. Since spatial resolution at higher eccentricity locations get noisier, peripheral presentation makes the localization of the whole display get noisier. Moreover, it has been overlooked that in the peripheral, position 2 in the first frame and position 3 in the second frame are actually crowded by the two elements on both sides. This crowding effect makes it very hard to precisely locate objects presented at these two locations, and thus it's very hard to apply the nearest neighbor rule to these objects. In the mean time, position 1 and 3 in the first frame, and position 2 and 4 in the second frame are not crowded. The relative velocity rule can be applied to them easily. Therefore, for 0 ms peripheral presented Ternus display, the output of the higher-level correspondence system is group motion perception.

On the other hand, for the lower-level transient detection correspondence system, with 0 ms ISI, motion energy at position 2 and 3 is clearly 0. Motion energy from position 1 to 4 is also very high. Therefore, the output from the lower-level motion system is biased to the element motion perception.

Unsurprisingly, it has been shown that at a 2 degree eccentricity, a 20 ms ISI display would generate about 60% group motion perception and 40% percent element motion perception (Breitmeyer & Ritter, 1986). These results give some initial evidence that the lower-level transient detection and higher-level correspondence systems are

complementary to each other. When they provide contradicting outputs, they will compete against each other and generate a final coherent perception.

In this experiment, I want to further explore how the two correspondence computation systems are involved in the perception of the Ternus display. More specifically, if the two systems are really complementary to each other, then adding difficulty to the computation of one system should bias the observer rely more on the other system. This hypothesis will be easier to test when the two systems are signaling opposite outputs, which is true for peripherally presented Ternus displays. I will use longer ISI conditions to decrease the amount of motion energy that can be used by the lower-level system (Adelson & Bergen, 1985). I will use crowding as an approach to decrease the spatial resolution of the higher-level correspondence system. I hypothesized that at shorter ISI and strong crowding conditions, transient luminance change in the spatio-temporal domain should be used more and thus produce more element motion perception. At longer ISI and no crowding conditions, position correspondence system should be favored and produce more group motion. It's unclear how the two systems would interact when both signals are very weak, i.e. in long ISI and strong crowding condition.

3.4.1 Method

Participants

21 undergraduate students at Johns Hopkins University participated in this experiment for course credit. All had normal or corrected-to-normal visual acuity.

Stimuli and procedure

Stimuli were presented on a Macintosh iMac computer with a refresh rate of 60 Hz. The viewing distance was approximately 60 cm so that the whole display subtended 39.4×24.8° of visual angle.

The Ternus display consisted of two frames of three red vertical lines (0.1° × 1°). In both frames, the three lines were separated by 1°. Across the two frames, the left most element shifted 1° such that the right two positions in the first frame and the left two positions in the second frame were overlapping (see **Figure 3.8** for an illustration of the display).

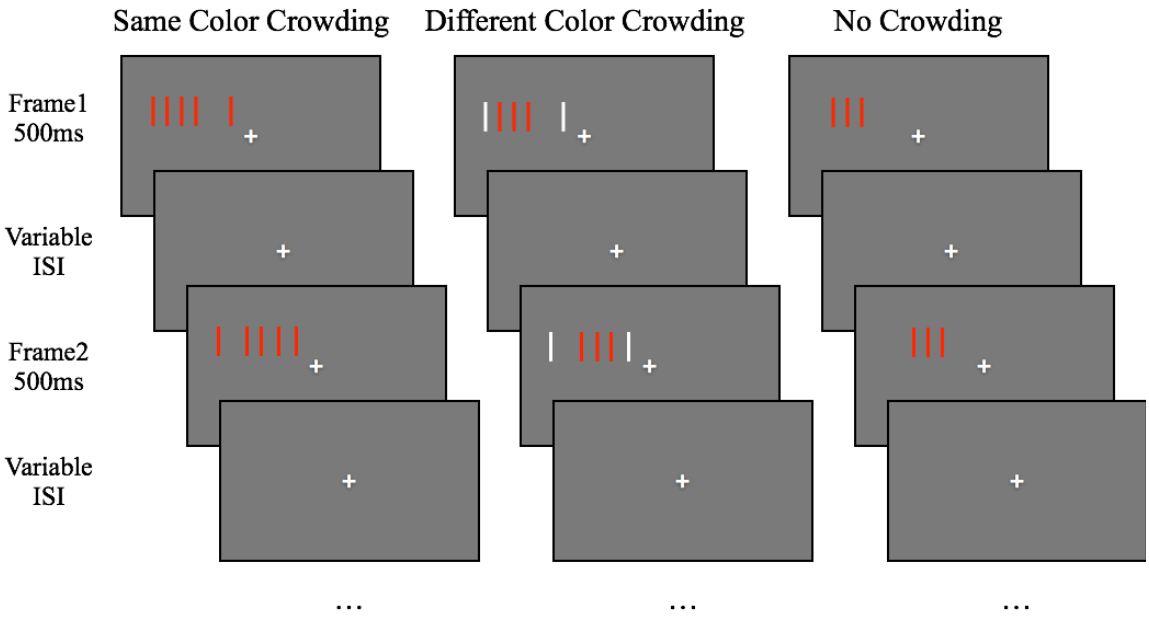


Figure 3.8 Experimental procedures of Experiment 6. Ternus display was presented in two eccentricity conditions and three crowding conditions with variable ISIs between the two frames. In each trial, the two frames were presented for six times alternatively.

The Ternus display was presented at two eccentricity conditions. In the low eccentricity condition, the center of the furthest element was presented 3.6° horizontally and 0.7° vertically away from the central fixation cross. In the high eccentricity condition,

the center of the furthest element was presented 8.3° horizontally and 3.1 vertically away from the central fixation cross.

In the no color crowding condition, the two frames were alternatively presented for six cycles. Each frame lasted 500 ms, with variable ISIs between the two frames. In the current experiment, 5 different ISIs were selected: 0 ms, 50 ms, 100 ms, 200 ms, and 501 ms. In the same color crowding condition, all procedures were the same except that two crowding bars were presented. The two bars had the same color as the Ternus display objects (red) and were located 0.5° to the left and right side of the whole display (**Figure 3.8**). The crowding bars remained on the screen during the variable ISIs. In the different color crowding condition, all procedures were the same to the same color crowding condition, except that the crowding colors were white. Different color bars were supposed to produce weaker crowding effect than same color bars (Whitney & Levi, 2011).

In sum, there were a total of 30 different conditions (2 eccentricity levels × 5 ISI levels × 3 crowding levels). Each participant completed 8 trials for each condition, leading to a total of 240 trials. All trials were presented in randomized orders. At the beginning of the whole experiment, a typical Ternus display was first presented to the participants. The experimenter introduced the two types of perception, element motion and group motion to the participants. The participants were also induced to see the two perceptions. After they said they understand the difference between the two perceptions, the main experiment started. During each trial, they were told to fixate at a central white cross and press one of two response keys to indicate whether the peripheral presented Ternus display induced more element motion or group motion. Since the Ternus display

was played for six cycles in each trial, they were told to make response for the dominant perception for the majority of the time.

I again used a secondary digit report task to encourage fixation at the central cross. At a random point between the second and the fifth cycle of each trial, a single digit (between 1 and 9) replaced the fixation cross for 100 ms. At the end of each trial, participants first reported the majority perception of the Ternus display, then entered the digit they saw.

3.4.2 Results

Data from one participant were excluded from further analysis due to very low overall digit report accuracy (22.9% versus 81.7% as the lowest accuracy for all other participants). I further excluded all trials with an inaccurate digit report of the remaining 20 participant, which amounted to 7.3% of trials.

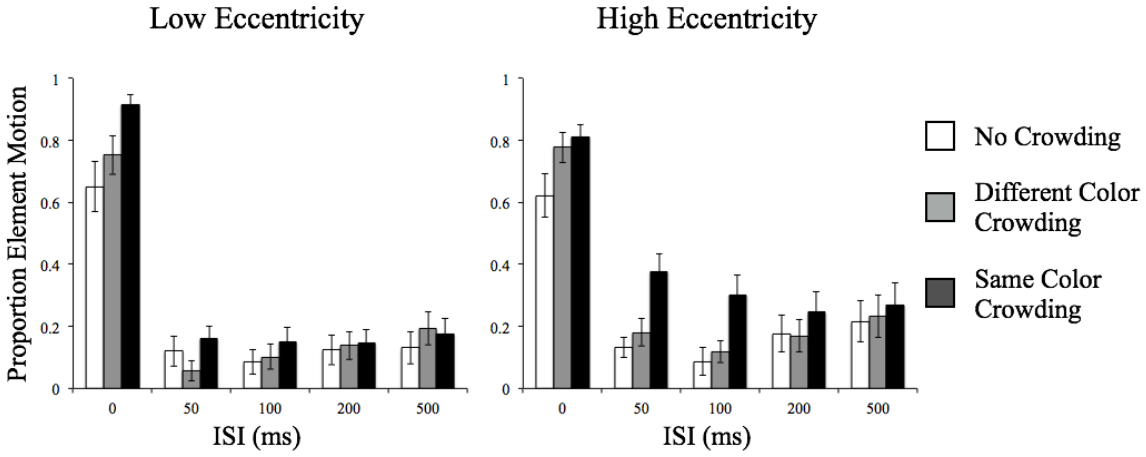


Figure 3.9 The proportion of element motion perception as a function of ISI, Crowding condition, and Eccentricity.

The proportion of trials in which participants reported seeing element motion under different conditions is plotted in **Figure 3.9**. A value of 0.8 in this graph shows that under that condition, participants on average reported 80% of the trials as element motion,

and 20% of the trials as group motion. A two (Eccentricity) by three (Crowding) by five (ISI) three-way repeated measure ANOVA showed that there was a significant main effect of Eccentricity ($F (1, 19) = 6.02$, $p = 0.024$, $\eta_p^2 = 0.24$), a significant main effect of Crowding, ($F (2, 38) = 13.2$, $p < 0.001$, $\eta_p^2 = 0.41$), and a significant main effect of ISI, ($F (4, 76) = 67.0$, $p<0.001$, $\eta_p^2 = 0.78$). The three two-way interactions were also significant: Eccentricity × Crowding, $F (2, 38) = 3.91$, $p =0.029$, $\eta_p^2 = 0.17$; Eccentricity × ISI, $F (4, 76) = 5.31$, $p = 0.01$, $\eta_p^2 = 0.22$; Crowding × ISI, $F (8, 152) = 4.2$, $p=0.004$, $\eta_p^2 = 0.18$. Critically, there was a significant three-way interaction: $F (8, 152) = 2.1$, $p = 0.039$, $\eta_p^2 = 0.10$ Note that whenever the sphericity assumption was violated, a Greenhouse-Geisser correction was applied.

Post-hoc contrast analysis suggested that the effect of higher eccentricity's role in increasing the proportion of group motion report is mainly observed in the 0 ms ISI condition ($F (1, 76) = 5.69$, $p < 0.05$ after Scheffé correction). No contrasts of the interaction between Eccentricity and crowding reached significant level. However, the effect seems to be driven by weaker crowding effect in the low eccentricity conditions.

To make the data more interpretable, I split the data by the eccentricity conditions, and ran two separate three (Crowding) by five (ISI) two-way repeated measure ANOVAs on the low eccentricity and high eccentricity data separately. For the low eccentricity condition, there was a significant main effect of Crowding ($F (2, 38) = 7.321$, $p=0.002$, $\eta_p^2 = 0.28$), a significant main effect of ISI ($F (4, 76) = 78.1$, $p<0.001$, $\eta_p^2 = 0.81$), and a significant interaction between the two factors ($F (8, 152) = 4.3$, $p=0.005$, $\eta_p^2 = 0.18$). For the high eccentricity condition, there was a significant main effect of Crowding ($F (2, 38) = 13.9$, $p<0.001$, $\eta_p^2 = 0.42$), a significant main effect of ISI

($F(4, 76) = 44.8$, $p<0.001$, $\eta_p^2=0.70$), and a significant interaction between the two factors ($F(8, 152) = 2.5$, $p=0.031$, $\eta_p^2= 0.12$). It's clear for both low and high eccentricity conditions, people reported more group motion with longer ISIs, and in general they reported more element motion under crowding conditions.

To more systematically analyze the three-way interaction effects, I ran 10 independent simple main effect tests on the effect of crowding conditions at each specific combination of Eccentricity and ISI. This will help us know all else being equal, whether people showed different bias to report element or group motion under different crowding conditions. With a Scheffé correction for multiple comparisons, we found that there were significant simple main effect of Crowding for the following conditions: 0 ms ISI in the low eccentricity condition, and 0 ms, 50 ms, 100 ms ISI in the high eccentricity condition. It's clear in the graph that under these conditions, external crowding stimuli significantly increase the probability to report element motion. However, at the longer ISI conditions under both low and high Eccentricity, people reported more group motion whenever the display was crowded or not.

3.4.3 Discussion

The current experiment first replicated some previous findings of the Ternus display. Both longer ISIs and peripheral presentation of the stimuli induced more group motion perception (Breitmeyer & Ritter, 1986; Petersik, 2006). Moreover, it could also solve some controversy in the literature. Unlike Breitmeyer and Ritter (1986), Petersik (2009) showed that the percentage of group motion did not change systematically as eccentricity change. However, in the data analysis, responses from different ISI conditions were combined. The current results suggested that the effect of eccentricity is

mainly observed in 0 ms ISI conditions. Therefore, similar effect could have been observed in the data of Petersik (2009) if it were analyzed based on different ISI conditions.

The most important result of the current experiment is the significant three-way interaction between Crowding, Eccentricity, and ISI. Crowding the whole display significantly biased participants to report more element motion, and same color crowding has a stronger effect than different color crowding. More importantly, this effect was only observed in relatively short ISI conditions. In the low eccentricity condition, the crowding effect was significant only in the 0 ms ISI conditions. In the high eccentricity condition, the crowding effect could be observed at two slightly longer ISI conditions, but did not reach significant levels for ISI longer than 100ms.

These results are in general consistent with our predictions: when crowding adds a lot of noise to the higher-level correspondence computation, the system was more biased to the output of the lower-level transient detection computation. When longer ISI weakens the transient signal of the lower-level computation, the system was more biased to the output of the higher-level position correspondence system. Finally, when both crowding and long ISI were present, the final output depends on the relative strength of the signals for the two systems.

The dynamic interactions of the two systems can be easily seen in the non-0ms ISI conditions. In the no crowding condition, all of the non-0 ms ISI conditions led to a majority of group motion perception. This could be explained by the fact that the transient detection system could only provide a weaker signal than the higher-level position comparison system. While it's still not very hard for the higher-level system to

compare the two outer positions in each frame, the lower-level system can only detect very weak motion signal from position 1 to position 4. When crowding was added, it should bias the system to rely more on the output from the lower-level system. However, this is highly dependent on the relative strength of the two types of signals. Crowding effects are higher in higher eccentricity conditions (Bouma, 1970, Levi, 2008; Pelli & Tillman, 2008). Therefore, in the higher eccentricity condition, the correspondence system receives noisier signals, which led the system rely more on the lower-level system. We thus observed an increase in element motion report under the same color crowding conditions. However, when the lower-level transient signal became weaker in the longer ISI conditions (200ms and 500ms), the higher-level system started to regain its role even when crowding was added. These results suggested that although both signals are weak in crowding long ISI conditions, probably the higher-level system can receive a less noisy signal than the lower-level system, and thus perception was biased to the output of the higher-level system.

In sum, the current study showed that different perceptions of the Ternus display were the result of the interaction between the lower-level transient detection system and higher-level position correspondence system. The two systems are complementary to each other such that the system with a stronger signal would bias the final perception more in accordance with its output. When presented in relatively peripheral locations, the element motion perception of the Ternus display relied more on the transient detectors, and the group motion perception relied more on the higher-level object-based correspondence computation.

3.5 General discussion

In this chapter, I first presented a brief case study of a patient, JKI, who suffered widespread brain damage, including extensive bilateral parietal lesions.

The patient was found to have two associated deficits largely restricted to the left visual field. First, he could not accurately localize the position of an object, although he could almost always report the object's shape and color. Second, the patient was impaired in judging the direction of slowly, but continuously (modally) moving targets, without impairment for faster moving targets (or for any motion in the right visual field). Broadly, JKI's motion perception deficit minimally demonstrates that higher-level correspondence computation might be necessary for the perception of slow smooth motion stimuli.

I am unaware of any reports of similar slow motion difficulty in either a patient study, or via experimental manipulation. There is, however, the well-known case study of LM, a patient with 'motion blindness', that is the inability to perceive motion (Zihl, von Cramon, & Mai, 1983; Zihl et al., 1991). In contrast to JKI, under many circumstances, LM is unable to detect or perceive motion as it unfolds (and the speeds tested in the relevant studies tended to be faster than any of the speeds we investigated here). But she can make directional judgments relatively accurately. She appears to achieve this through input from higher-level mechanisms, comparing an object's current and remembered positions. In her words, "First the target is completely at rest. Then it suddenly jumps upwards and downwards" (referring to vertical motion; Zihl et al., 1991, pp. 2244). Thus LM appears not to possess the phenomenological experience of stimulus motion, although she does appear to possess the higher-level, position- and object-based correspondence mechanisms that seem unavailable to JKI's left visual field.

It could be that the difference between LM and JKI is not a dissociation between phenomenal experience and higher-level judgment mechanisms, but between phenomenal experience and lower-level signal integration. JKI's neurological condition may simply have made him *more* susceptible than normal to the inherent difficulty of perceiving slow motion with low-level detectors, with the implication that his motion deficit is unrelated to his spatial representation deficit. This concern was especially salient given some damage in JKI in MT/MST.

Although distinguishing between these possibilities would be difficult in the case of JKI, I sought to investigate the association between explicit position representation and slow motion perception in the case of healthy observers. Specifically, I investigated the possibility that normal perception of slow, smooth motion might rely heavily on higher-level correspondence computations that depend on representation of a target's position and compare the positions across time.

To test this hypothesis, I induced a slow-motion perception deficit in visually normal observers via crowding. The crowding inducers in these experiments were static. Their purpose was to make it difficult for an observer to resolve the current position of a moving target within the flanked region, as crowding is known to impair the resolution of object individuation spatially and to limit the ability to allocate object-based attention (Intrilligator & Cavanagh, 2001; Pelli & Tillman, 2008; Whitney & Levi, 2011). The result suggested for normal people, the higher-level correspondence system is needed to correctly judge the direction of slowly moving stimuli.

Finally, I took the advantage of peripheral presented Ternus displays. In such displays, the output of the higher-level correspondence system would prefer group

motion perception, while the output of the lower-level transient signal detection system would prefer element motion perception. I thus used crowding and longer ISI to increase the noise in the two systems respectively. The results suggested that the two correspondence systems are complementary to each other. The relative signal strength determines the final output. Motion perception in the Ternus display is often consistent with the system that has relative less noise in the input signals.

In general, one may wonder why the visual system has evolved to employ two distinct correspondence computation algorithms for motion perception. Providing a comprehensive answer to this question is beyond the scope of the current data. But the results across the three experiments together do suggest that the systems are to some degree complementary rather than redundant. The lower-level system efficiently detects quick luminance transients, but such a system will always depend on the sensitivity setting of its constituent detectors. It appears that those detectors may have been set with temporal and spatial parameters that produce false negatives with respect to continuous but slow motion (perhaps in an effort to otherwise reduce false positives.) Fortunately, higher-level mechanisms can compensate by relying on representation of token position.

This perspective is consistent with additional ways in which the two systems may be complementary, even compensatory. The lower-level system, without any explicit representation of object positions, suffers from the aperture problem (Adelson & Movshon, 1982). Therefore, feature tracking is needed to solve motion correspondence between features (Ullman, 1979). It may also be that lower-level mechanisms are best for detecting motion direction, with coarse grained direction sensitivity (Hildreth, 1984), whereas higher-level mechanisms can disambiguate motion direction when necessary

(Shimojo, Silverman, Nakayama, 1989). And more recently, it has been suggested that binocular feature tracking is necessary to overcome the inverse problem of local binocular three-dimensional motion perception (Lages, 2013; Pierce et al., 2013; Lages & Heron, 2010).

In sum, these experiments suggest that the lower-level and the higher-level correspondence systems are both involved in motion perception. Depending on the nature of the signal in the stimuli, the two systems work complimentary to each other. The higher-level system is needed to perceive slow but smooth motion stimuli, while the lower-level system is needed to perceive the element motion of the Ternus display. Under the framework of correspondence computations, motion perception is the output of different algorithms that are balanced based on the available input signal to each system.

Chapter 4: Eye-movements and correspondence computation in multiple object tracking (MOT) task

Multiple object tracking (MOT; Pylyshyn & Storm, 1988) has been one of the most productive paradigms for understanding the nature of human visual attention. In a typical MOT trial, participants were asked to track a set of moving targets among a larger set of objects (See **Figure 4.1A** for an illustration of the procedure of one trial). The difficulty of this task could be easily manipulated by many basic factors, such as target load, object speed, and tracking duration. Human observers tend to perform well with relatively moderate speeds when asked to track between three and five targets, but not more (Alvarez & Franconeri, 2007; Scholl, 2009).

The MOT paradigm was first invented to study the properties of visual attention. Along this line, previous studies often tried to explain the capacity limits observed in MOT tasks by the limited amount of available attentional resources, with either a fixed-slot or flexible nature. However, the computations, especially correspondence computations, involved in the MOT task have been largely ignored. In many previous theories, it is not clear what kind of representations attention deals with, and how attention makes the tracking ability possible.

I propose that the performance limits could be explained by a different kind of explanation that focuses on the inherent processes and computations rather than resources that could be used up. The distinction between the two kinds of explanations is clear in terms of computer programs. For example, there are multiple ways to sort an array of numbers into ascending order. With the same computer settings, one sorting algorithm (e.g. Quicksort) could be much faster than another (e.g. Bubble sort). Therefore, the

speed limit can only be explained if we start to appreciate the difference in the algorithms and computations used by different sorting methods. Similarly, to understand the cause of the observed performance limits in visual attention tasks like MOT, one also needs to look at the computational level.

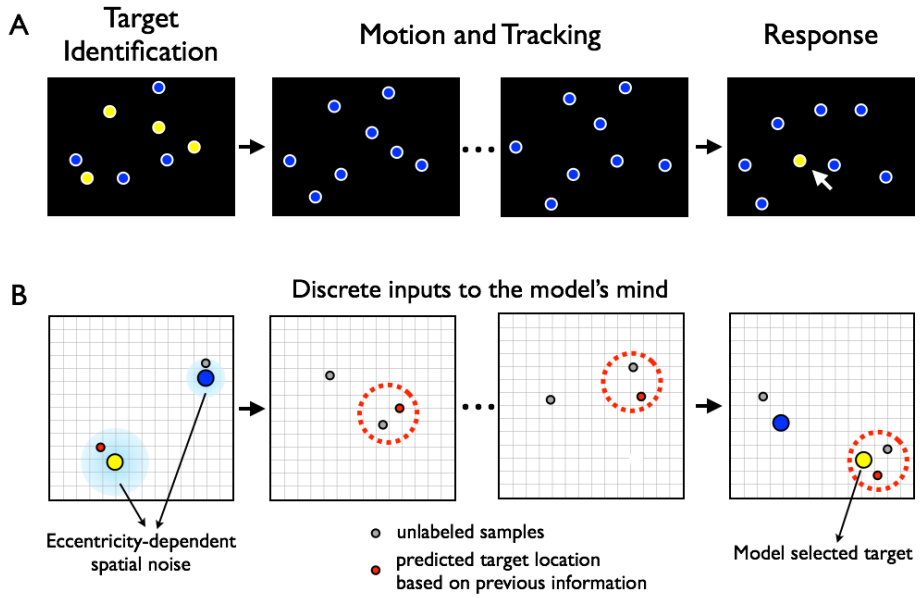


Figure 4.1 A: An Illustration of the procedure of a typical MOT trial. At the beginning of the trial, target objects were shown by changing to a different color. Then all objects changed back to the same color and start to move randomly across the display. Participants were asked to track the original targets. At the end of the trial, participants use a mouse cursor to clicked at object that they think are targets. **B:** A schematic illustration of key properties of the proposed computational model. During the target identification period, the model labeled samples from target objects as targets. During the motion period, the model samples the visual display discretely, and represent each object's spatial location with eccentricity-dependent noise. It uses probabilistic analysis and nearest neighbor rule to correspond observations from the current time sample to predicted target locations based on previous information. Note that we only illustrated observed samples from one target and one distractor here. The real model gets noisy observations from all objects and makes correspondence analysis for the targets.

Before we start to understand the algorithms and computations involved in a typical MOT task, it's very important for us to figure out the kind of inputs the system can get. This is because any algorithm should be suited for the kind of input and output representations of the system (Marr, 1982). Attention gets the input from lower-level perception and pass on the information to some higher-level processes. Therefore, we need to first figure out the kind of constraints that have been imposed by the perceptual system. It's worth noting that these physical constraints from the perceptual system could finally limit human tracking performance, but they are at different levels than the limits caused by algorithms and computations used during tracking.

We hypothesized that at least two physiological aspects that can affect the inputs to the attentional tracking system, and thus further affect tracking performance. First, it has been widely shown that the human visual system cannot sample the world consecutively. Human perception might completely rely on discrete processing epochs (VanRullen & Koch, 2003), and the 10 Hz rhythm in the occipital lobe is closely related to visual perception (VanRullen & Macdonald, 2012). Landau and Fries (2012) showed that human selective attention is modulated by a rhythm at about 4 Hz. Using a modified MOT task with a circular trajectory, Holcombe & Chen (2013) showed that people were able to track a single target at a sampling rate of 7 Hz, and this rate would decrease to 4 Hz when they are tracking two targets. Therefore, during MOT, it is highly possible that the observers are dealing with discrete rather than continuous visual inputs. It will be meaningful to test this hypothesis, and understand tracking computations based on inputs with a discrete nature.

Second, in addition to the limit in the temporal domain, there is also perceptual limit in the spatial domain. Visual receptors are densely packed in the fovea, with increasingly fewer amount in the periphery (Young, 1971). This organization is duplicated in the brain, where visual areas represent foveal inputs more precisely than peripheral inputs (Carrasco & Frieder, 1997). This means that people are always uncertain about the spatial locations that are in their peripheral visual field. Since it's impossible to fixate at all targets at the same time, almost all target locations in an MOT task can only be represented noisily. And this uncertainty in spatial representation should play an important role in limiting tracking performance.

Besides the inherent physical limits of the visual system, other cognitive processes could also change the quality of visual inputs. Eye-movement pattern, or the location the person is fixating at, is a critical factor that determines the uncertainty of spatial representations at different locations. Tracking errors can easily happen when a distractor is too close to a target (Bae & Flombaum, 2012). A more precise spatial representation is needed when a target has a close encounter distractor. On the other hand, a noisy spatial representation might be good enough to track a target that is far a way from all of the distractors. Therefore, whether a target and a nontarget will be confused is a function of where they are relative to the fovea, and eye-movement patterns clearly change the eccentricity levels of different objects. Many studies have explored the spontaneously generated eye-movement patterns during MOT tasks (Fehd & Seiffert, 2008, 2010; Zelinsky & Neider, 2008). However, there is still a lack of research investigating where people should look, and how much tracking performance will be affected by change in eye-movement patterns. The current study also aimed to study

these important questions about eye-movements during MOT, and build up computational explanations based on the findings of eye-movements.

While it is very important to understand the nature of inputs used during tracking, our ultimate goal is to know what kind of computations a system can do with these exact inputs, and how well the system can perform with these computations. More importantly, what kind of computations could well simulate human tracking performance with these visual inputs. The correspondence computation of current received signals and those from previous moments is the core process of the MOT task. I hypothesized that human visual computation is probabilistic. Noisy knowledge of object positions produces confusions between targets and nontargets probabilistically (Ma & Huang, 2009; Vul et al., 2009; Franconeri, Jonathan, & Scimeca, 2010; Bae & Flombaum, 2012;). People are trying to apply nearest neighbor rules to make correspondence analysis for objects across tracking frames. We developed a computational model that can make best guesses on current object locations based on its noisy spatial representations and probabilistic inference (See **Figure 4.1B** for a schematic illustration of important model properties). I hypothesized that the observed capacity limit in MOT will arise naturally after taking this probabilistic computation into consideration.

In this chapter, I planned to answer four major questions. First, what is the temporal rate that observers can sample a visual display during tracking? Second, what is the spatial resolution at different eccentricity locations during tracking? Third, how eye-movement can affect visual inputs and thus affect tracking performance? Will different eye-movement pattern bring big difference in tracking performance? Finally, what are the computations people might use to track multiple moving targets at the same time? After

considering the inherent physiological and computational limits, do we need a form of cognitive resources to explain human limits in tracking ability? To study these research questions, we combined evidence from behavioral experiments and computational simulations. We first used two behavioral experiments to figure out the range of human spatial resolution and temporal sampling rate during tracking. We then systematically investigated the influence of eye-movements to human tracking performance. We also developed an algorithm that is supposed to find the best fixation locations at each frame of a tracking trial. Finally, we implemented a modified Kalman Filter model to investigate the possible computations used by human observers during tracking. The model received discrete, noisy inputs from moving targets and nontargets that were corrupted by noise dependent on the distance of the input from current fixation. Surprisingly, the model accounted for most of the variance in human performance as a function of tracking load and speed, and it also explained differences between individuals very well.

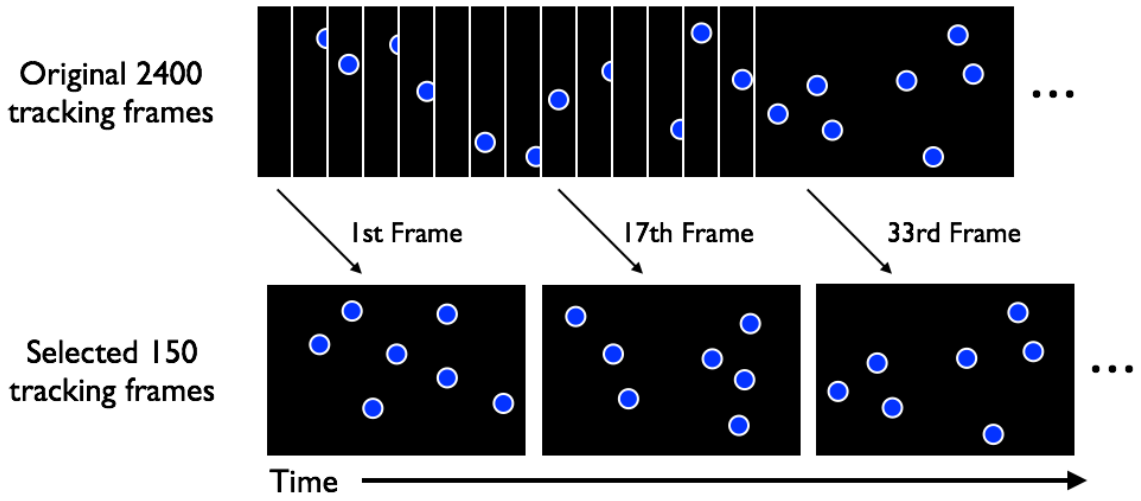
4.1 Experiment 7: People have a limited temporal sampling rate during MOT

In this experiment, I wanted to get an estimate of the range of human observers' temporal resolution when performing an MOT task. This is a question that has been either ignored, or over simplified in previous literature. For example, one previous study on computational models of MOT has simple used the refreshing rate of computer monitor as human's temporal resolution to sample the visual display (60 Hz, Vul et al., 2009). However, this is actually an empirical question that could be tested by experiments.

We took the following hypothesis: if during tracking, people could sample the visual display at a really high refreshing rate, e.g. 60 Hz, then tracking performance should be about the same when the same amount of information is presented at different frame durations. This is because people should always be able to catch up with the pace of the presentation rate, and process the information contained in successive presented frames, no matter each frame was there for 16.7ms or 100ms. However, if people can only sample the visual world at a relatively low rate, e.g. 20 Hz, then they would miss to process two thirds of the tracking frames if each frame is there only for 16.7ms. This is because with a 20 Hz sampling rate, human participants would need 50ms to process one frame. Three frames would have been presented during that period and thus the visual information could not be fully processed.

In this circumstance, lowering the presentation rate from 60 Hz to 20 Hz should enable participants to fully process each tracking frame and significantly improve performance.

To directly test this hypothesis, for each pre-generated tracking trial, we presented the same 150 tracking frames with different frame durations. (**Figure 4.2**). The amount of information we provided to the observers was exactly the same across different frame duration conditions. However, the required sampling rates to fully process the information are different. A 16.7ms frame duration will require a 60Hz sampling rate, and a 100ms frame duration will only require a 100Hz sampling rate. If people were able to sample the visual world at a very high rate, we should observer equally high performance across different frame rate conditions. A decrease of performance at a certain frame duration would suggest that human's sampling cycle is longer than that duration, and thus sampling rate is smaller than the corresponding required sampling rate.



Frame Duration (ms)	Total Tracking Duration (s)	Required sampling rate (Hz)
16.7	2.5	60
33.3	5	30
50.0	7.5	20
66.7	10	15
83.3	12.5	12
100	15	10

Figure 4.2 Design and different conditions of Experiment 8. A tracking trajectory was first generated with 2400 frames. Then, one of every 16 frames was selected to compose the 150 to-be-presented frames. In different conditions, each frame was presented for different frame durations. The total tracking duration and required human sampling rate to successfully process each frame are listed to the right of each frame duration.

4.1.1 Method

Trajectories

All of the stimuli in this and all following experiments were generated with MATLAB Version 2014a (MathWorks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). We first randomly generated 120 trajectories that have two targets and five distractors. Each trial began with seven disks (.43° radius) and two of

them were selected as targets. No object could begin a trial closer than 1.25° to any other object (measured center to center). Each object began moving at a fixed speed (1 pixel per frame) in a randomly determined direction. When objects collided with the boundary of the display, with the center point of the display, or came within 1.25° of one another, they deflected according to Newtonian principles. Each trial originally contained 2400 frames and would last 40s if presented regularly on a 60 Hz monitor (16.7 ms per frame). The object moved 1 pixel per frame. The display as a whole subtended 27° x 20°.

For each trajectory, we selected the 1st, 17th, 33rd, …, and 2385th frames as the 150 to-be-presented frames (**Figure 4.2**). I decided not to directly generate these 150 frames because by originally setting the speed to 1 pixel per frame, the dynamic between objects would be smoother. Each trajectory was played in one of the six possible frame duration conditions: 16.7, 33.3, 50.0, 66.7, 83.3, and 100 ms. The corresponding required human sampling rates are: 60, 30, 20, 15, 12 and 10 Hz. The same amount of information was presented to the participants across different conditions. The only difference was the amount of time the participants were allowed to process these information. All participants did the same 120 tracking trials, 20 in each condition.

The assignment of which trajectory to which condition was completely randomized for each participant, with the only constrain that no participant would see the same tracking trial in different conditions.

Participants

17 Johns Hopkins University undergraduates participated in this experiment. All had normal or corrected-to-normal visual acuity. The protocol of this and all following behavioral experiments was approved by the Homewood Institutional Review Board of

Johns Hopkins University. Data from one participant was excluded in the analysis due to extremely low overall tracking accuracy (36.3% versus an average of 86.4% of all other participants and a chance level at 28.6%).

Apparatus and Procedure

Stimuli were presented on a Macintosh iMAC computer with a refresh rate of 60 Hz. The viewing distance was approximately 60 cm so that the whole monitor subtended 39.43° × 24.76° of visual angle. Stimuli were presented with MATLAB and the Psychophysics toolbox (Brainard, 1997; Pelli, 1997).

All stimuli were presented in a black square subtending 27° x 20°. Each trial started with seven discs (.43° radius) along with a white fixation cross (0.47°× 0.47°) in the center. After .5s, two discs turned yellow for 1.5s, indicating that they were the targets. Finally, all discs turned blue again. After another .5s, all of the discs moved, following pre-selected trajectories (see above) for 2.5s to 15s. At the end of the motion, participants were prompted to click on the discs that they thought were the targets.

The participants were told nothing about the nature of the trajectories, and their only task was to click out the two targets at the end of each trial. Each participant completed a total of 120 trials.

4.1.2 Result

Participants' tracking performance at each frame duration condition is plotted in **Figure 4.3.** A repeated measure one way ANOVA showed that there was a significant main effect of presentation rate condition ($F(5, 75) = 58.7$, $p<0.001$). Post-hoc contrast analysis (with Greenhouse-Geisser and Scheffé correction) showed that, there was a significant difference between the 33.3 ms and 16.7 ms conditions ($F(1, 15)=82.1$,

p<0.05). Besides, there was a significant difference between 33.3 ms and the average of 50 to 100 ms conditions ($F(1, 15)=20.6$, $p<0.05$). There were no significant differences among the four longer tracking durations (all pairwise $p > 0.05$). These results suggested that 33 ms is not enough for people to finish sampling the information of one tracking frame, and the human sampling rate during tracking is probably below 20 Hz. However, since we were not sure whether each frame in the 150 frames was critical to the tracking task, failing to process a couple of them might still lead to similar performance. Therefore, we cannot make any further conclusion about human's true sampling rate in the tracking task. Though people might have different sampling rate for different number of targets, the range we've got here could still provide a good approximation to what people are using in a typical MOT task. Therefore, in the following studies, we will use sampling rates under 20 Hz as the human temporal rate to sample visual inputs during tracking. More specifically, to cover the basic range we found in this experiment, we will focus on two sampling rates, 12 Hz and 20 Hz.

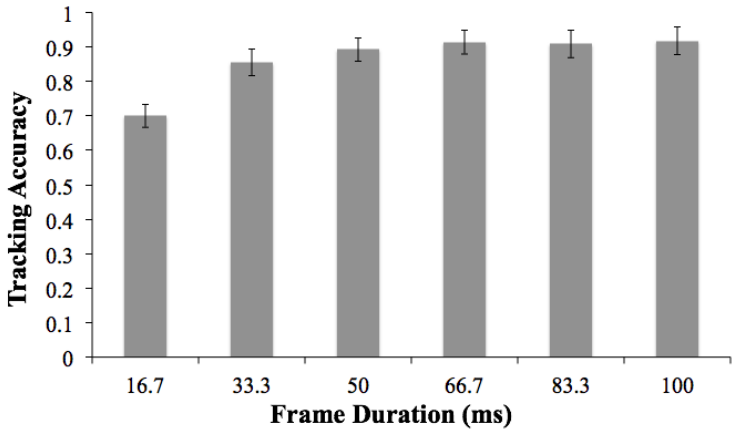


Figure 4.3 Average tracking accuracy at different frame duration conditions in Experiment 8. Error bars reflect standard errors of the mean.

4.1.3 Discussion

The results of the current experiment mainly suggested two things. First, the limits in the human sampling rate of visual information could affect tracking performance. Second, under a typical MOT setting, human visual sampling rate is between 10 to 20 Hz. In the next experiment, we will move on to explore the relationship between human limits in the spatial representation.

4.2 Experiment 8: Measuring people's noisy spatial resolution

The current experiment aimed to estimate people's spatial resolution across the visual field. Its result will be further used in later experiments to discuss the realtionship between spatial resolution and MOT.

People only have noisy representations of spatial locations, especially for locations that are in the periphery. This is mainly due to the fact that the central visual field is represented by larger brain areas than the peripheral visual field (Daniel & Whitteridge, 1961). Moreover, the cortical magnification factor M (mm of cortex per 1 degree of visual angle at different retinal locations) is positively correlated to the density of ganglion cells on the retina (Drasdo, 1977). It has been shown that for the superior visual field, the relationship between M and eccentricity in degree of visual angle (E) could be summarized by the following function: $M(E) = M_0(1 + 0.42E + 0.00012E^3)^{-1}$ (Rovamo & Virsu, 1979; Virsu & Rovamo, 1979; Carrasco & Frieder, 1997). Rovamo and Virsu (1979) have also suggested that the spatial resolution at a certain eccentricity location could be predicted accurately by multiplying a constant to the value of M at that location. Therefore, the noise of the spatial representation at eccentricity E, which is the

inverse value of the spatial resolution, or the standard deviation of a two-dimensional Gaussian distribution centered at eccentricity E, can be calculated by the following formula: $\sigma_z^2(x) = c(1+0.42E)$ (Vul et al., 2009, note that the $0.00012E^3$ term has been left out here because its value is much smaller than the other two terms). In this formula, all parameters are based on empirically measured results, except for the unknown value c (which reflects the spatial noise at fovea location). Vul et al. (2009) has used c=0.08 in their computational model of MOT to simulate human uncertainty in spatial representation. However, the authors did not give detailed explanation in why they chose this value. In fact, the spatial resolution near the fovea area hasn't been systematically studied before. Therefore, we designed this experiment to get an empirical estimate of the spatial standard deviation at fovea. Since any computations must be based on a certain type of input, knowing the exact noise in visual inputs will further help us understand computations involved in MOT.

We used the method of constant stimuli to infer people's spatial resolution at the fovea. Critically, we assumed that whenever a participant saw an object and infer its spatial location, he was actually getting a random sample from a two dimensional normal distribution centered at the object's true location with a standard deviation σ. When two stimuli were presented sequentially, the participant would get two samples from the two distributions independently. When the participant was asked to make a judgment on whether the two objects were presented at the same or different locations ("No Move" vs. "Move"), he would apply an internal criterion such that if the distance between the two samples exceeded the criteria, he would make a "Move" response. We further made the assumptions that participants would apply the same criterion across different conditions,

and that σ would take the same value at different locations that are slightly away from the fovea. In the experiment, participants saw two discs presented sequentially at two different locations that are separated by a range of distances. We were thus able to infer back the participants' σ and criteria based on their psychophysical function of reporting "Move" given different distance,

4.2.1 Method

Participants

Three Johns Hopkins University undergraduate students participated in this experiment. All had normal or corrected-to-normal visual acuity.

Apparatus and Stimuli

Stimuli were presented on a LCD monitor with a refreshing rate of 60 Hz, controlled by a Mac mini (Apple Inc., Cupertino, CA). The viewing distance was approximately 55 cm so that the whole display subtended 40.4° x 30.7° of visual angle. All stimuli were generated and presented with MATLAB Version 2014a (MathWorks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

Procedure

Each trial started with the presentation of a blue disc (.6 radius) that is similar to the objects we used in the MOT tasks. The disc could appear on any where within a 38° × 18° area centered on the screen. The participants were told to make a saccade to the target disc as soon as possible, such that the target disc would be presented at the fovea of the participant. The disc would remain on the screen for 1s, providing enough time for the participant to fixate at the disc location. Then, after a 0.5s blank interval, another identical disc would be presented, but with its location shifted either to the left or the

right side of the location of the first disc. There were ten levels of shifting distance, varied from 0.02 degree to 0.2 degree. The participants were asked to press one of two keys to indicate whether they thought the second dot was presented at the same ("No Move") or different ("Move") locations to the first disc. The second disc disappeared as soon as the participant made a key response (**Figure 4.4**).

Each participant completed 48 trials for each shifting distance condition, such that the three participants together provided 144 trials for each shifting distance condition for further analysis.

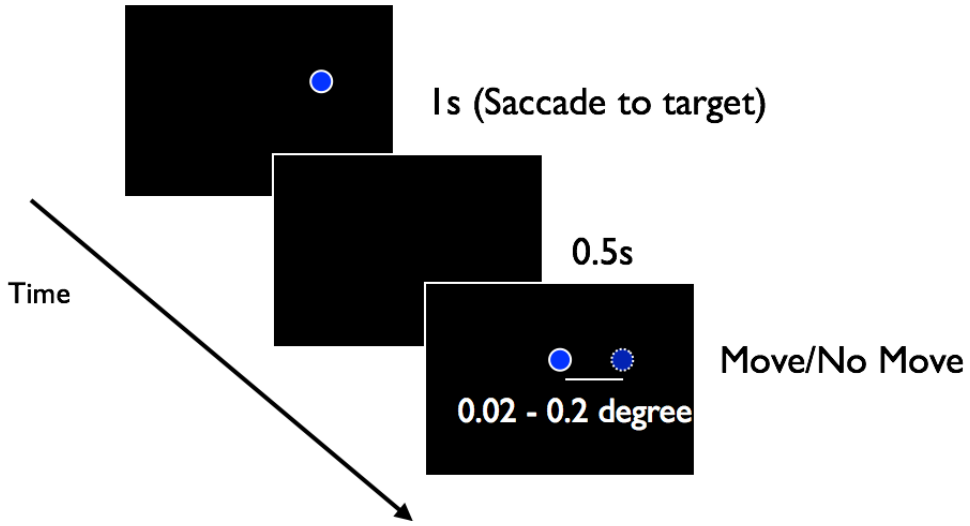


Figure 4.4 Procedure of Experiment 8.

4.2.2 Result

In all of the following analysis, we combined the responses from all participants together. We first calculated participants' proportion of "Move" responses given different shifting distances, and these were plotted in **Figure 4.5**. In accordance to one's intuitive expectation, the probability of reporting "Move" increased as the true moving distance increased.

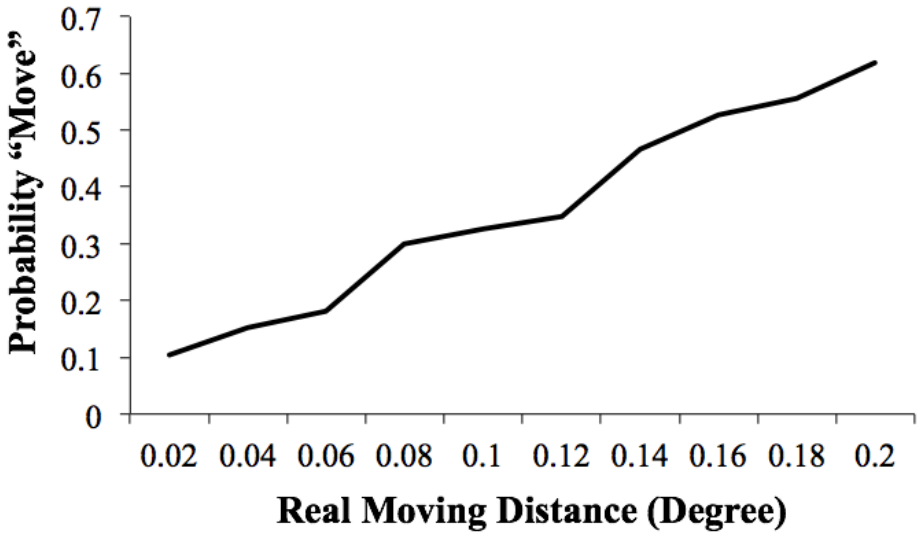


Figure 4.5 The probability of "Move" responses given the real moving distance of the second disc.

We then needed to solve a problem with two unknown parameters: the standard deviation of spatial representation at fovea σ_{fovea}, and the criteria that people would call a moving distance as 'Move'. To fully use all of the information to solve the problem, we simulated a model and used a least square estimation to infer the σ_{fovea} and criterion values that were driving this observed behavior data.

For a trial that had a true shifting distance of x, and for a given pair of σ_{fovea} and criterion values (denoted as $\sigma_{fovea\text{-}i}$ and $criterion_{\text{-}i}$ here), the model would get two random samples, one from N (0, $\sigma_{fovea\text{-}i}$) and the other one from N (x, $\sigma_{fovea\text{-}i}$). Then the model would compute the distance between the two samples. If this distance is larger than $criterion_{\text{-}i}$, the model would provide a "Move" response. Otherwise, the model would provide a "No Move" response. For each candidate pair of $\sigma_{fovea\text{-}i}$ and $criterion_{\text{-}i}$ values,

we ran the model 144 times (the same number of trials with human participants) at each shifting distance conditions and calculated the corresponding proportion of "Move" responses. We then get the squared difference between human responses and model responses and use this value as a goodness-of-fit measure of the current $\sigma_{fovea\text{-}i}$ and $criterion_i$ pair. We searched through a space where σ_{fovea} could take any value between 0.02° and 0.34°, and criterion could take any value between 0.02° and 0.34°. Each simulation would give us the pair of σ_{fovea} and criterion values that lead to the least squared errors. We then run this simulation 100 times, and the average best fitting σ_{fovea} and criterion values were 0.082° and 0.176°.

4.2.3 Discussion

In the current experiment, I have combined the method of constant stimuli and model simulation to measure human's spatial resolution at the center of visual field. Using a least square estimation, we found the best fitting standard deviation at fovea location is 0.082°. This value is very close to the 0.08 used by Vul et al. (2009). The result of the current experiment provided a foundation for our future experiments. It confirmed the idea that representation of spatial locations is noisy, and it's an important factor that should be considered in any formal theory of tracking mechanisms. In our following discussions, we will include this spatial uncertainty as an innegligible constrain on human tracking ability. We will use 0.08 as the standard deviation of spatial represenations near foveated locations.

4.3 Experiment 9: The influence of eye-movements to MOT performance

In natural vision, people often make frequent eye-movements to collect and integrate information of the visual world (Hayhoe, 2000). With multiple targets to track, one critical problem for human observers is where to put the fovea. Since one component of MOT is to represent the locations of each target, the spatial resolution at a target location should play a critical role in determining the successful tracking of that target. While it has been proposed that qualitatively, the requirement of fixation at the display center does not affect tracking performance (Scholl & Pylyshyn, 1999), there is a lack of direct experiments to further support this idea. Many studies have used people's eye-movement patterns during tracking to make inference on the strategies they used. People tended to use a target-centroid looking strategy, at least for trials with three or less targets (Fehd & Seiffert, 2008; Zelinsky & Neider, 2008). When people were asked to track as many as four targets, they tended to look more at individual targets, or use a strategy that shifts between target centroid and individual targets.

Fehd and Seiffert (2010) showed that tracking performance was higher when participants were switching fixation locations between target and target centroid, than when they were switching among targets. This result provided some initial evidence that eye-movement pattern during tracking is critical in determining tracking performance. However, it is still unclear how much individual difference in tracking performance could be caused by different eye-movement patterns. Besides, eye-movements have often been studied with a limited number of objects (e.g. eight). The field has been underestimating the influence of eye gaze locations in determining tracking performance.

In this experiment, we systematically investigated the relationship between eye-movements and tracking performance. We hypothesized that if eye-movement pattern is critical to tracking, then people adopted different eye-movement patterns should have different tracking performance. We thus collected participants' spontaneously generated eye-movement pattern when they were performing typical tracking tasks. We then explored whether there are any systematic relationships between the eye-movement patterns and tracking performance. This would help us to evaluate the importance of eye-movement to tracking limits.

4.3.1 Method

Trajectories

Each MOT trial began with six to sixteen disks (.6° radius) and half of them were targets. No object could begin a trial closer than 1.75° to any other object (measured center to center). Each object began moving in a fixed speed in a randomly determined direction. When objects collided with the boundary of the display or with the center point of the display, they deflected according to Newtonian principles. When objects came within 1.75° of one another, they bounced off each other in a Newtonian way. Each trial lasted 10s.

We included four different tracking speeds: 2.8°/s, 5.6°/s, 8.4°/s and 11.2°/s. We generated 50 trials for each combination of target load and speed, and this led to a total of 1200 trials. We then randomly selected five trials from each combination of target load and speed. We tested all participants with the same 120 trajectories.

Participants

We recruited 10 Johns Hopkins University undergraduate and graduate students to participate in this experiment. All had normal or corrected-to-normal visual acuity.

Procedure

Stimuli were presented on a LCD monitor with a refreshing rate of 60 Hz, controlled by a Mac mini (Apple Inc., Cupertino, CA). The viewing distance was approximately 55 cm so that the tracking area subtended 40.4° x 30.7° of visual angle.

Participants' eye positions were collected simultaneously using an EyeLink 1000 desk-mounted eye-tracker system (SR Research, ON), with a sampling rate of 500 Hz. Participants were instructed to freely move their eyes in the way they think can help them track better.

Each trial started with six to 16 discs (.6° radius) along with a white fixation cross (0.6°× 0.6°) in the center. After .5s, half of the discs turned yellow for 1.5s, indicating that they were the targets. Finally, all discs turned blue again. After another .5s, all of the discs moved, following pre-selected trajectories (see above) for 10s. At the end of the motion, participants were prompted to click on the discs that they thought were the targets (**Figure 4.6**).

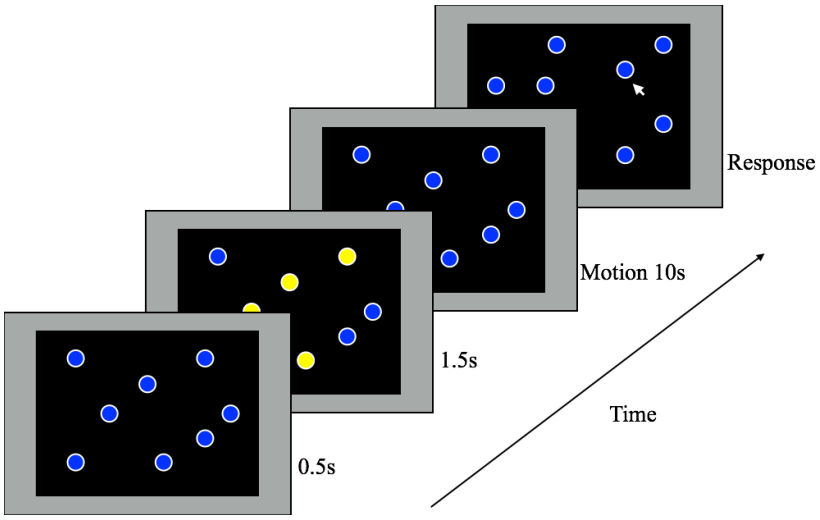


Figure 4.6 The procedure of a typical MOT trial used in Experiment 9 and 10.

Eye-tracking data analysis

With a sampling rate of 500 Hz, the eye-tracker constantly collected 5000 gaze location data points for each 10s MOT trial. This led to a massive amount of data for further analysis. Therefore, we pre-processed the data in the following way. First, we categorized all of the eye-tracking data into 'fixations', 'saccades' and 'blinks' using the eye-tracker's online parser with the following criteria: saccades were determined when eye velocity was greater than 30 degree/s, acceleration was greater than 8000 degree/s^2, and saccadic motion was larger than 0.15 degree; blinks were determined when the pupil is missing; all of the other moments were categorized as fixations. All of the consecutive 'fixation' moments were averaged to get a single pair of x and y coordinates for the whole period. Second, with a 60 Hz monitor, each MOT trial contained 600 motion frames. We thus match the time of all of the fixation periods with the 600 frames. Any motion frame that fell between the ending time of the i^{th} fixation period and the ending time of the i+1th fixation period was assigned the average gaze location of the i+1th fixation period. After this pre-analysis, we've got 600 pairs of eye gaze coordinates that corresponded to the

600 motion frames. We used these coordinates for all of the model simulations and data analysis.

4.3.2 Result

Participants' average tracking accuracies at different target load and speed conditions are plotted in **Figure 4.7**. A 6 (target load) × 4 (speed) repeated measure ANOVA shows that there was a significant main effect of target load ($F(5, 45) = 128$, $p<0.001$), a significant main effect of speed ($F(3, 27) = 196$, $p < 0.001$), and a significant interaction between target load and speed ($F(15, 135) = .96$, $p<0.001$). These results are consistent with previous findings that increase in either target load and speed will decrease tracking performance (e.g. Pylyshyn & Storm, 1988; Oksama & Hyönä, 2004; Liu et al., 2005; Alvarez & Franconeri, 2007).

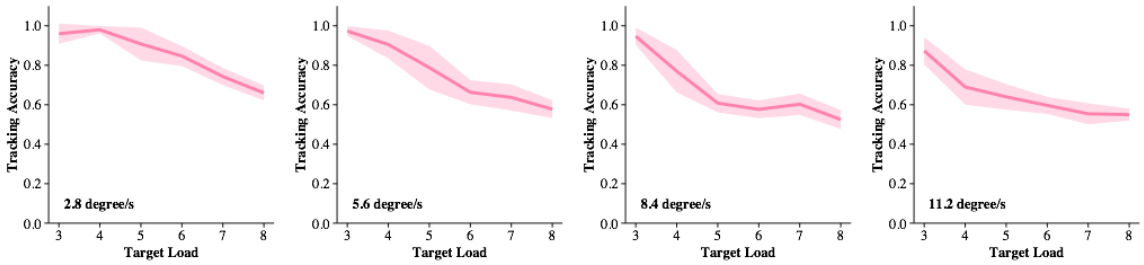


Figure 4.7 Average human tracking accuracy at different target load and speed conditions in Experiment 9. The thickness of the lines reflects 95% confidence interval of the mean.

We then used two analyses to explore the relationship between participants' eye-movement pattern and their tracking performance.

In the first analysis, we sorted the 10 participants' tracking performance for each of the 120 MOT trials separately. For one specific trial, we then obtained the ranks of different participants. Since there are only several possible tracking accuracy values for each trial (e.g. 0, 0.33, 0.66 and 1 in a three-target trial), it is highly possible that two or

more participants could get the same rank, including the best or worst rank. If eye-movement pattern is critical in determining tracking performance, then participants with similar tracking performance should have used similar eye-movement strategies, while people that differed a lot in tracking performance should have used different eye-movement patterns. Based on this hypothesis, we first grouped participants into three possible pairs. A best-best pair consisted of two participants that both had the highest tracking performance in one specific trial. A worst-worst pair consisted of two participants that both had the lowest tracking performance in one specific trial. Finally, a best-worst pair consisted of one participant that did the best, and one participant did the worst in one specific trial. Across all 120 tracking trials, we've got 919 best-best pairs, 214 worst-worst pairs, and 754 best-worst pairs.

For each pair, we calculated the correlation between the two participants' eye-movement coordinates (combining x and y values into one vector) across the 600 frames in one trial. We also calculated the average distance between the two participants' eye-movement coordinates. We used both the correlations and the distance to serve as the measurement of the similarity between two participants' eye-movement patterns. More similar eye-movement patterns will have higher correlation and lower average distance.

The averaged correlation and distance for the three pair conditions are shown in **Figure 4.8A** and **4.8B**. One-way ANOVA suggested that there is a significant main effect of pair type for both the correlation ($F(2, 1884) = 39.5$, $p<0.001$) and the distance ($F(2, 1884) = 89.3$, $p<0.001$). Pairwise post-hoc contrast analysis showed that the correlation within the best-best pair is the highest, followed by the best-worst pair, and then the worst-worst pair. Similarly, the distance within the best-best pair is the smallest, followed

by the best worst pair, and then the worst-worst pair (all ps<0.001 after Bonferroni correction). These results suggested that good trackers often adopt similar eye-movement patterns, while people that cannot track very well may adopt different eye-movement strategies.

In the second analysis, we aimed to explore the general relationship between eye-movement patterns and tracking performance. We used the total amount of eye-movement distance as the measurement of eye-movement pattern. We selected this measurement since the total distance of eye-movement is calculated based on the eye-gaze locations across all time points in a tracking trial, and thus should reflect a person's general eye-movement strategy. For each participant, we calculated his or her average eye-movement distance across all trials, and correlate it with the same person's average tracking accuracy (**Figure 4.9**). There is a significant negative correlation between the eye-movement distance and tracking accuracy (r=-0.78, p=0.007), with more eye-movement distance predicting lower tracking performance.

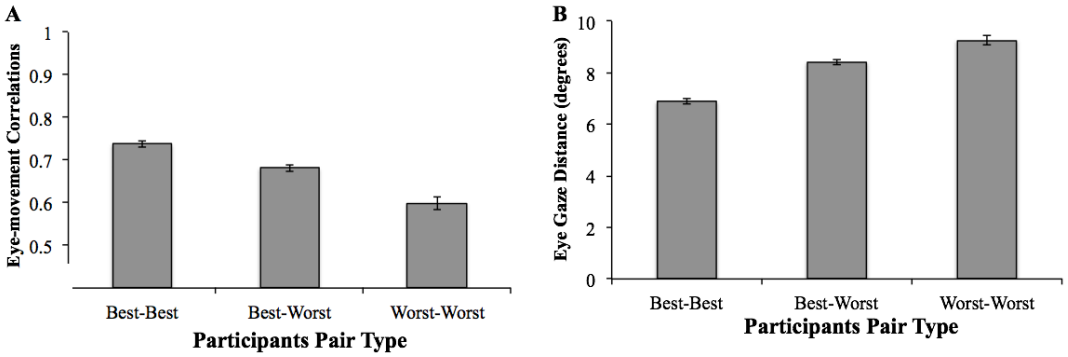


Figure 4.8 A: Eye-movement correlation for different types of participants pairs in Experiment 9. B: Eye gaze location distance for different types of participants pairs in Experiment 9.

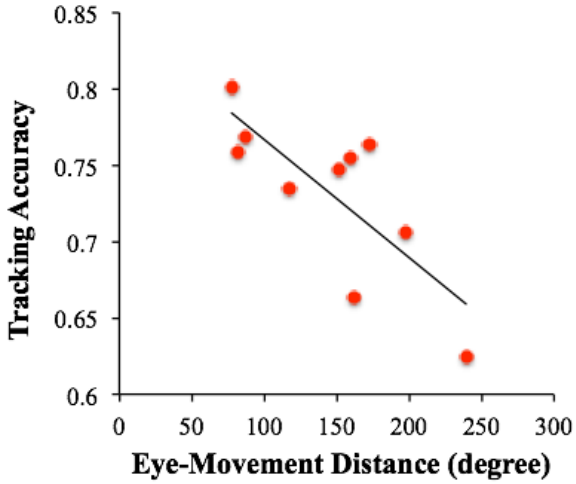


Figure 4.9 The correlation between participants' tracking accuracy and their average total eye-movement distance in each trial.

4.3.3 Discussion

In this experiment, we've shown that people's eye-movement patterns during tracking are critical to their final tracking accuracy. Shorter eye-movement distance will lead to higher tracking accuracy. Besides, people with higher tracking accuracy seem to use similar eye-movement strategies, while people with lower accuracy could use different strategies. This result suggested that for each trial, there might be one (type of) eye-movement strategy that can in general lead to higher tracking performance.

In sum, the evidence from this behavior experiment confirmed the hypothesis that fixation selection an important factor in tracking multiple objects. Putting the fovea at different locations may lead to completely different tracking performance of the same trial. Therefore, as described below, our computational simulation of the MOT task will adopt human eye-movement pattern to simulate each individual participant's tracking performance.

4.4 Computational Experiment 10: A probabilistic computational model that can simulate human performance in MOT

In the above experiments, I've shown that the inherent limits in the human visual system are critical to determining the limits of human tracking ability. However, these factors could not provide complete answers to the following questions: why people can track some targets successfully and why they will fail to track in some other cases. In other words, we still don't understand the computations underlying human tracking abilities. In the current experiment, we tried to explore the kind of algorithms and computations used in MOT, given the constrains of noise and uncertainty in the inputs to the system. We hypothesized that human visual computation, and thus tracking process is probabilistic. Noisy knowledge of object positions produces confusions between targets and nontargets probabilistically (Bae & Flombaum, 2012). Based on this hypothesis, we implemented a computational model that performed the MOT task. This model used a modified Kalman Filter algorithm to keep track of multiple targets simultaneously. It applied nearest neighbor rule to make correspondence targets identities across tracking frames. Based on the results in the previous experiments, this model adopted eye fixations obtained from human observers tested on the same task in the lab. The model received inputs from moving targets and nontargets that were corrupted by noise dependent on the distance of the input from current fixation. The model also sampled the visual world discretely, in the rate range we figured out in Experiment 7. The inclusion of these physiological constraints made our model a lot different from previous computational models of MOT (Vul et al., 2009; Ma & Huang 2009). We predicted that

the model could capture variability in human performance without any form of external added resource limit.

4.4.1 General model framework

The Kalman Filter is a Bayesian model that tracks stochastic linear dynamical systems observed through noisy sensors. It operates on a stream of noisy input data to produce a statistically optimal moment-by-moment estimate of the underlying system state (here, positions and velocities). The Kalman Filter is a recursive estimator, which means that it makes successive predictions and then corrects these predictions in light of new observations. This amounts to a form of feedback control: the model predicts the system state at some time and then obtains feedback in the form of (noisy) measurements. Accordingly, equations for the Kalman Filter can be classified as either prediction equations or measurement equations. Prediction equations use probabilistic beliefs about the current state and recent past to obtain prior estimates for the immediate future. Measurement/observation equations are responsible for the feedback—for using new measurements to obtain posterior state estimates that may differ from the priors. Details of this process can be found in Zhong et al., 2014 and other widely available sources deriving and describing the Kalman Filter more exhaustively (Kalman, 1960; Welch & Bishop, 2006; Yilmaz, Javed, & Shah, 2006; Murphy, 2012).

Observation

At a given moment in time t, $\mathbf{z}_t^m$ denotes the m^{th} observation of position—including vertical and horizontal coordinates— from an object in the display, m=1,…,N_A. An observation is derived from the object's true state, denoted, $\mathbf{l}_t^m$, the position and velocity of object m at time t, as follows:

$$\mathbf{z}_t^m = \mathbf{H}_k \mathbf{l}_t^m + \mathbf{r}_t^m \quad (1)$$

Here, $\mathbf{H}_k$ is the mapping matrix, which maps the true state space into the observed space, such that $\mathbf{H}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$. $\mathbf{r}_t^m$ is observation noise, assumed to be zero-mean Gaussian white noise with measurement noise covariance $\mathbf{R}_t^m = \sigma_z^2 \mathbf{I}_2$. With a fixed value for σ_z^2 one obtains a model with uniform spatial resolution throughout the visual field. In our models, we utilized eccentricity dependent σ_z^2 values given by $\sigma_z^2(E) = c(1+0.42E)$ (Rovamo & Virsu, 1979; Carrasco & Frieder, 1997), where E is the distance between an item's true position and the observer's eye fixation location. c was a constant set to 0.08 based on our empirically measured value in Experiment 8 and previous studies (Vul et al., 2009).

Based on the result in Experiment 7, the model also had a limited temporal resolution of 12 Hz or 20 Hz, which means that the model only got samples from the objects 12/20 times per second.

Inference

Given an observation that has been assigned to a particular target m at time t, the model estimates a posterior for the target's current state, denoted $\hat{\mathbf{l}}_t^m$. This estimate is obtained by the weighted combination of a prior position estimate assigned to time t, $\tilde{\mathbf{l}}_t^m$, and the observation:

$$\hat{\mathbf{l}}_t^m = \tilde{\mathbf{l}}_t^m + \mathbf{K}_t^m(\mathbf{z}_t^k - \mathbf{H}_t \tilde{\mathbf{l}}_t^m) \quad (2)$$

Note that the index for the observation, $\mathbf{z}_t^k$, need not be the same as that of the object (i.e., the observer may have associated the wrong measurement with a target being

tracked). $\mathbf{K}_t^m$ is the weight matrix, also called the 'Kalman gain', which determines the relative weight of the prior and the current observation in determining the posterior estimate. The value for $\mathbf{K}_t^m$ is selected to minimize the error covariance in the posterior, denoted $\hat{\mathbf{P}}_t^m$ (Jacobs, 1993). Similarly, $\tilde{\mathbf{P}}_t^m$ denotes the error covariance in the prior at time *t*. $\mathbf{K}_t^m$ and $\hat{\mathbf{P}}_t^m$ are thus obtained via the following pair of equations:

$$\mathbf{K}_t^m = \tilde{\mathbf{P}}_t^m \mathbf{H}_t^T (\mathbf{H}_t \tilde{\mathbf{P}}_t^m \mathbf{H}_t^T + \mathbf{R}_t^m)^{-1} \quad (3)$$

$$\hat{\mathbf{P}}_t^m = (\mathbf{I}_2 - \mathbf{K}_t^m \mathbf{H}_t) \tilde{\mathbf{P}}_t^m \quad (4)$$

Prediction.

To understand how the model obtains prior estimates, consider time *t*+1. The expected position of the object should depend on basic motion kinematics, projecting forward from the posterior estimated at time *t,* $\hat{\mathbf{I}}_t^m$.

$$\tilde{\mathbf{I}}_{t+1}^m = \mathbf{M}_t \hat{\mathbf{I}}_t^m \quad (5)$$

Here, the 4x4 matrix $\mathbf{M}_t$ in the difference equation is the state-transition matrix based on basic motion. Such that $\mathbf{M}_t = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

When the model makes a prediction about an object's future position, it also projects forward an expected error covariance in the prior which is denoted as $\tilde{\mathbf{P}}_{t+1}^m$, to be used at time *t*+1. This estimate is derived from the difference between the prior and the posterior position estimates at the previous time point (Bishop, 2006):

$$\tilde{\mathbf{P}}_{t+1}^m = \mathbf{M}_t \hat{\mathbf{P}}_t^m \mathbf{M}_t^T + [\hat{\mathbf{I}}_t^m - \tilde{\mathbf{I}}_t^m][\hat{\mathbf{I}}_t^m - \tilde{\mathbf{I}}_t^m]^T \quad (6)$$

Correspondence

In typical computer vision applications, the correspondence problem —the problem of linking measurements with objects being tracked— is not solved on a purely spatiotemporal basis. This is because only one object is tracked, or knowing the identity of an object is not important for the task, or because differences among objects in surface appearance (such as color or shape) can be utilized. In the MOT paradigm, however, the identities of multiple objects are important (at least at the level of the target vs. nontarget distinction), and perceptual differences other than position are not available to inform correspondence inferences. Our model address correspondences in the following way.

We denote $p(T_t^m = k)$ as the probability that the k^{th} observation at time t corresponds to target m. The model attempts to solve the correspondence by assuming that a new observation for target m will be drawn from a Gaussian distribution centered on the adjusted prior expectation about the position of m, $\mathbf{A}_t^m$. Thus:

$$p(T_t^m = k) = N(\mathbf{z}_t^n; \mathbf{A}_t^m), \quad 1 \le k \le N_A, 1 \le m \le N_T \tag{7}$$

This adjusted prior expectation is a mixture of pure priors and previous posterior positions. For time t+1, it denoted as $\mathbf{A}_{t+1}^m$, is obtained via the following equation:

$$\mathbf{A}_{t+1}^m = (1 - \beta_t^m)\tilde{\mathbf{I}}_{t+1}^m + \beta_t^m \hat{\mathbf{I}}_t^m \tag{8}$$

where β_t^m is the average of the main diagonal value in the Kalman gain matrix of target m, $\mathbf{K}_t^m$.

Assuming that the new observations are generated independently, and incorporating the principles of mutual exclusivity and exhaustive association for all

objects, the optimal correspondence can be obtained by maximizing the probability in Equation 9.

$$\{k_m \mid 1 \le m \le N_A\} = \underset{1 \le k_m \le N_A}{\operatorname{argmax}}\left[\prod_{1 \le m \le N_A} p(T_t^m = k_m)\right] \quad (9)$$

This is equivalent to minimizing the sum or product of the Mahalanobis distances (equivalently, the Euclidean distances) of new observations and the expected positions of the targets they are assigned to. There are other heuristic approaches to the correspondence problem, based on nearest-neighbor matching or specific validation regions (BarShalom et al., 2009; Murphy, 2012), that could be explored in future models.

4.4.2 Human and Model Testing Method

Trajectories

I used the same trajectories as in Experiment 9 to test both human participants and the model.

Participants

20 participants participated in the behavior part of this experiment. Each of them was asked to perform 120 trials of standard MOT tasks with varying target load (3-8 out of 6-16 total objects) and speed (2.8°/s, 5.6°/s, 8.4°/s and 11.2°/s) conditions. The 20 participants were further divided into two groups that have 10 participants. 10 participants in group1 each watched different 120 trajectories so that the groups as a whole completed a total of 1200 different trajectories. The other 10 participants in group2 were the same as in Experiment 9. As explained before, they watched exactly the same 120 MOT trajectories to allow for direct individual difference comparison.

Procedure

Human testing procedure and eye-movement analysis were exactly the same as in Experiment 9. The pre-processed eye-movements from each participant were further sent to the model of MOT. The model tracked the same trajectories as the human observers with the corresponding eye-movement patterns. The model performed each trial with each participant's eye-movement patterns 100 independent times, performing differently each time because of randomly generated measurement noise. In this way, each human observer had a corresponding simulated observer generated by the model. Tracking accuracy, calculated by the number of correctly selected targets out of the total number of targets, was used for further analysis.

4.4.3 Result

We observed very similar performance from human participants and the model (**Figure 4.10A** and **4.10B** for the 12 Hz model, **Figure 4.11A** and **4.11B** for the 20 Hz model). There were significant main effect of target load and speed condition on both human and model tracking performance (target load: $F(5,45)=114$ for human group1, $F(5,45)=172$ for 12 Hz simulated group1, $F(5,45)=114$, for 20 Hz simulated group1, $F(5,45)=128$ for human group 2, $F(5,45)=479$ for 12 Hz simulated group2, $F(5,45)=407$ for 20 Hz simulated group2; speed : $F(3,27)=126$ for human group1, $F(3,27)=167$ for 12 Hz simulated group1, $F(3,27)=126$ for 20 Hz simulated group1, $F(3,27)=196$ for human group2, $F(3,27)=548$ for 12 Hz simulated group2, $F(3,27)=27$ for 20 Hz simulated group2, all $ps<0.001$, with Greenhouse-Geisser correction if needed). For group1, the root-mean-square-deviation (RMSE) was 0.168 and 0.168 for the 12 and 20 Hz model respectively, and for group2, the RMSE was 0.170 and 0.165 respectively. These results

suggested that on average the model prediction deviated about 0.17 from the real human tracking accuracy. Overall, without fitting any parameters, the model was very successful in capturing average human tracking performance.

We then further investigated whether the model was successful in capturing the individual differences among different participants. For each specific combination of speed and target load, we calculated the average tracking accuracy for each participant and his/her corresponding simulated participant. After controlling the effect of target load and speed, partial correlation analysis suggested that there were significant positive correlations between human performance and corresponding simulated model performance (12 Hz group1: $r(236)=0.467$, $p<0.001$, 20 Hz group1: $r(236)=0.442$, $p<0.001$, 12 Hz group2: $r(236)=0.386$, $p<0.001$, 20 Hz group2: $r(236)= 0.470$, **Figure 4.12**). These results suggested that after adopting eye-movement patterns collected from individual observers, the model could capture the variability among human observers very well.

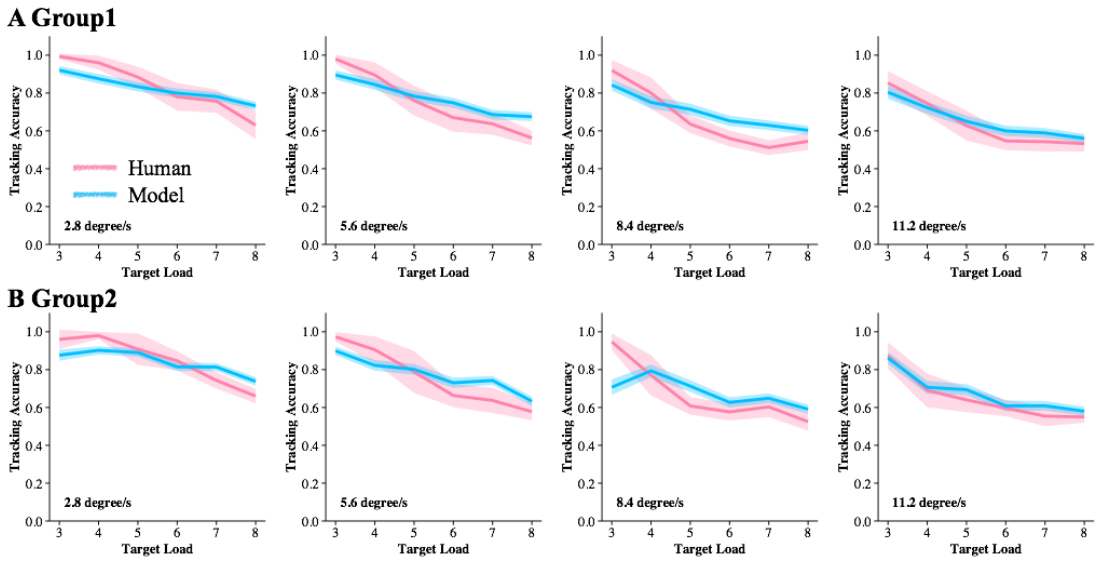


Figure 4.10 Average group 1 human participants (10 participants did different trials, panel **A**) and group 2 human participants (10 participants did the same 120 trials, panel **B**) tracking performance, together with simulated 12 Hz model tracking performance as a function of target load and speed. The thickness of the lines reflect 95% confidence intervals of the mean.

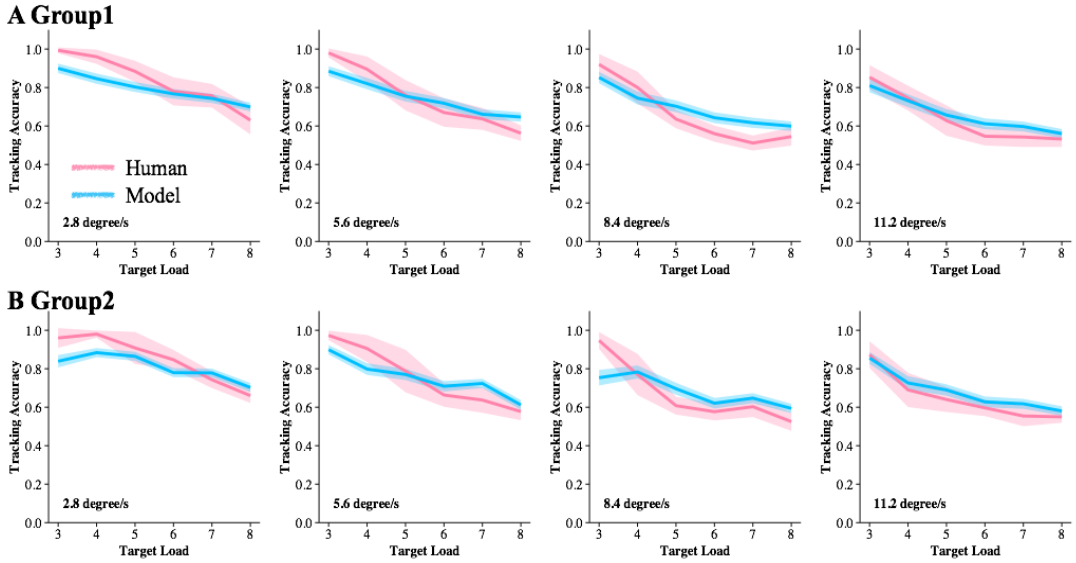


Figure 4.11 A. Average group 1 human participants (10 participants did different trials, panel **A**) and group 2 human participants (10 participants did the same 120 trials, panel **B**) tracking performance, together with simulated 20 Hz model tracking performance as a function of target load and speed.

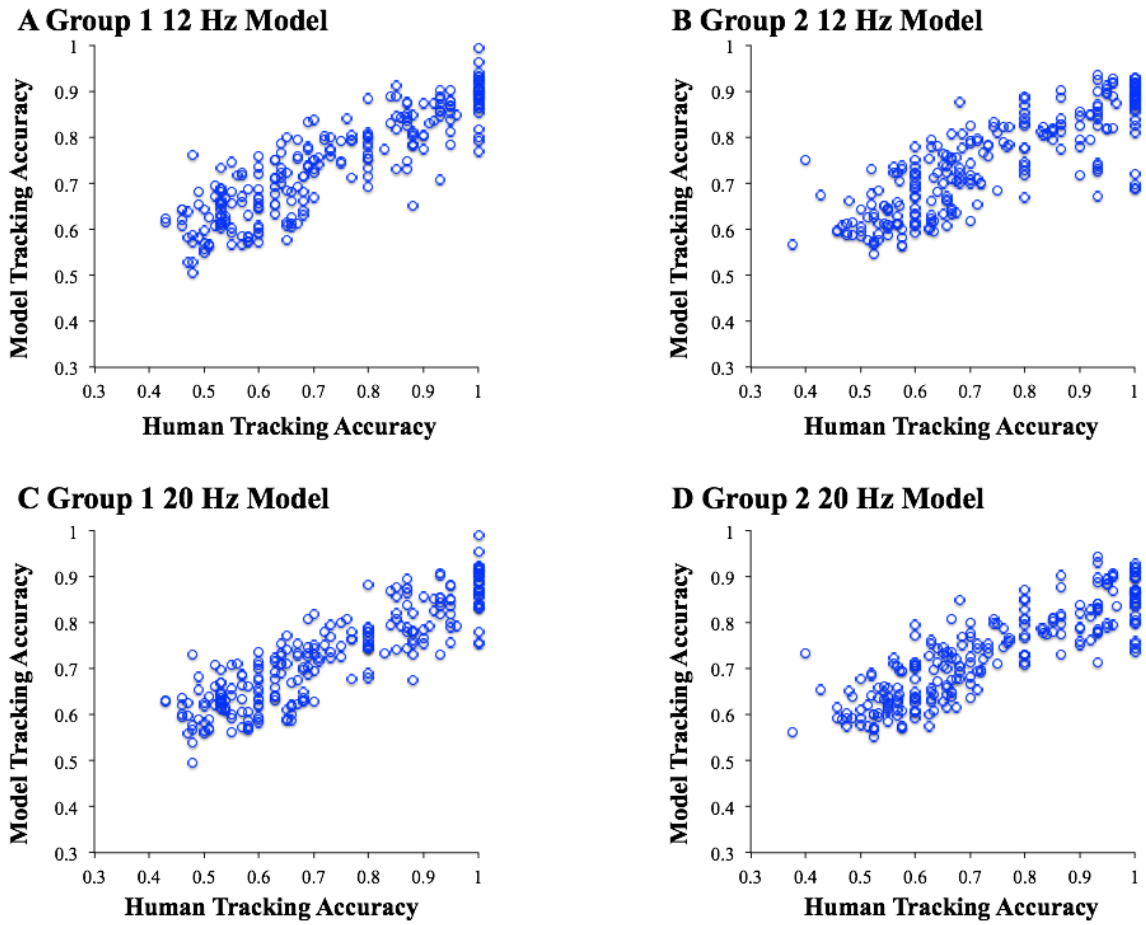


Figure 4.12 Correlations between human and model performance across different individuals. Each dot represents a participant at a particular target load and speed condition.

We also tested whether the model would make similar correct and incorrect responses as human observers. Since all group2 participants did exactly the same 120 trials, their data provided a great opportunity for us to answer this question. For each object in each trial, we first computed the probability that human have selected it as target. We then correlated this human selection probability with the probability that model had selected it as target. We computed this correlation score for targets and distractors separately. If the model often made similar choices as human observers, no matter correct or incorrect, we should observe significant positive correlations for both

targets and distractors. In fact, we did get significant positive correlations between human and model selection probability for both targets (12 Hz: $r(658)=0.461$, $p<0.001$, 20 Hz: $r(658)=0.496$, $p<0.001$) and distractors (12 Hz: $r(658)=0.372$, $p<0.001$, 20 Hz: $r(658)=0.415$, $p<0.001$, see **Figure 4.13**).

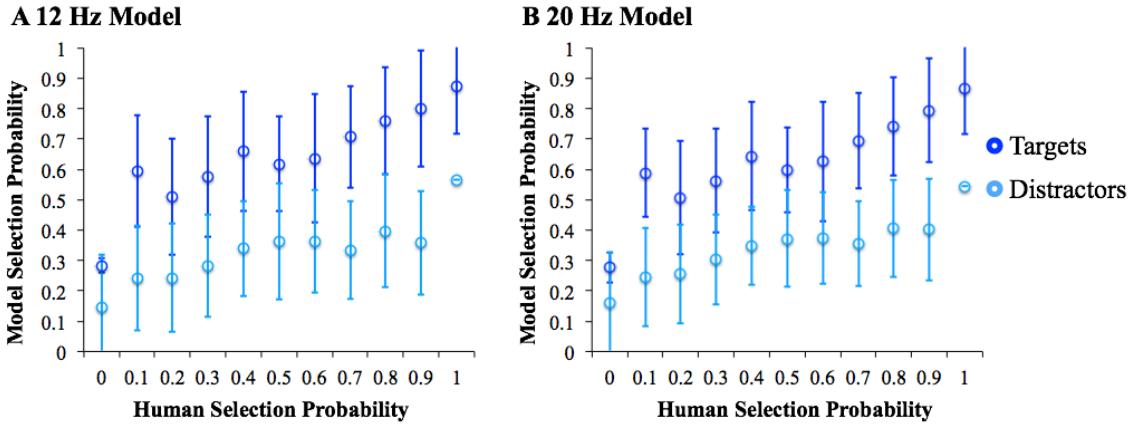


Figure 4.13 Correlation between human and model behavior in selecting a certain object as target. The x axis shows the probability that human observers have selected an object as target, the y axis shows the probability range that the model have selected the same object as target. Dark and light blue colors represent the selection probability for real targets and real distractors respectively.

4.4.4 Discussion

The current model could capture a great amount of variance in tracking performance, including the difference among different observers, among different trajectories, and even among different objects within a single trajectory. However, it has been suggested that formally adding constrain from a limited cognitive resource should explain human tracking behavior better. Besides, comparison among different models is critical in model evaluation. Therefore, in Experiment 10b, we ran another model that has constrains from a form of limited cognitive resources. This limited-resource model is the same as the resource-free model except for its target load-dependent spatial resolution.

We hoped to compare the model's goodness of fit between the two and infer whether an external cognitive resource is necessary in explaining human tracking limit.

4.5 Computational Experiment 10b: A probabilistic computational model of MOT with constraint from external cognitive resources

The results of Experiment 10 have suggested that a model considered probabilistic computations, eccentricity-dependent noisy, limited temporal resolution, and human eye-movement strategies can explain human performance in MOT tasks well. However, it has been a longstanding theory that externally imposed resources constrain human tracking ability (e.g. Alvarez and Franconeri, 2007; Vul et al., 2009). Therefore, we also explored this possibility in our proposed computational model of MOT. We formalized the cognitive resource by changing the value c of the spatial resolution function: $\sigma(E) = c(1 + 0.42E)$. It has been shown that the precision of spatial representation is highly dependent on the amount of available resource. With a fixed amount of cognitive resource, as the number of to-be-remembered locations (K) increased, the resource could be allocated to each location decreased, and thus the precision would also decrease (Alvarez & Franconeri, 2007; Ma & Huang, 2009; Vul et al., 2009; Holcombe & Chen, 2011). The relationship between the number of targets and the precision of spatial representation (P) could be described by a power low function ($P \propto (\frac{1}{K})^{0.74}$, Bays & Husain, 2008). Therefore, we re-calculated the value c for each target load condition with the following function: $c(K) = 0.0355(K^{0.74})$, with 0.08 as the value for target load of three, and 0.1654 as the value for target load of eight. Again, the limited-resource model adopted human eye-movement patterns and completed the

same trajectories as human participants. The limited-resource model was exactly the same as the resource-free model except for using different c values at different target load conditions.

4.5.1 Results

For the 12 Hz model, overall, there were significant main effect of target load and speed condition on model tracking performance (**Figure 4.14**, target load: $F(5,45)=480.5$, $p<0.001$ for simulated group1, $F(5,45)=832.8$, $p<0.001$ for simulated group2; speed : $F(3,27)=70.6$, $p<0.001$ for simulated group1, $F(3,27)=528.8$, $p<0.001$ for simulated group2). However, RMSE was larger for the limited-resource model than the resource-free model, though the t-tests did not reach significant level (group1: 0.180 vs. 0.168, $t(9)=1.32$, $p=0.22$; group2: 0.182 vs. 0.169, $t(9)=1.53$, $p=0.16$), suggesting that the limited-resource model deviated equally or more from the real human performance.

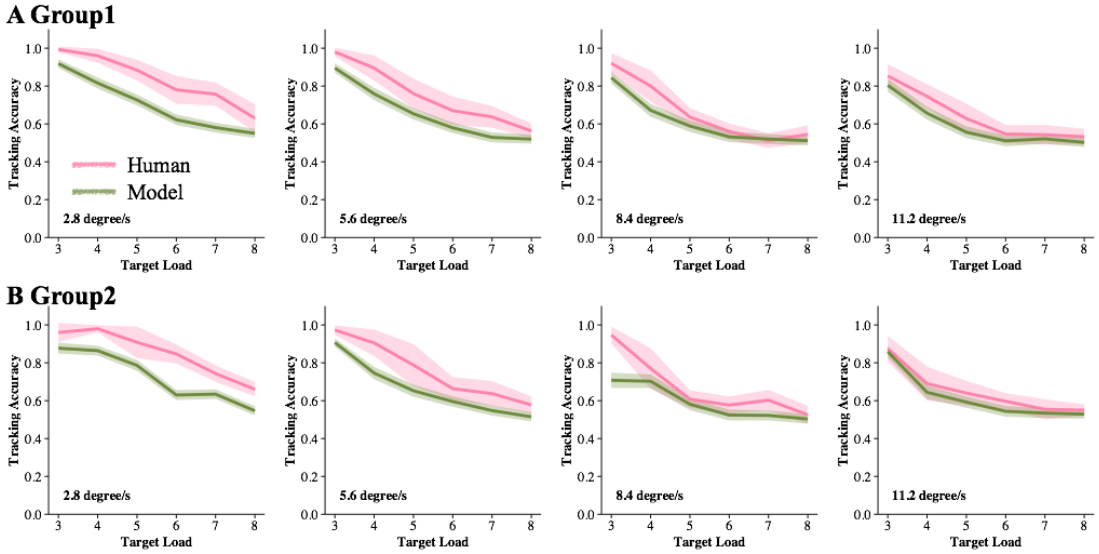


Figure 4.14 Average group 1 human participants (10 participants did different trials, panel **A**) and group 2 human participants (10 participants did the same 120 trials, panel **B**) tracking performance, together with simulated 12 Hz limited-resource model tracking performance as a function of target load and speed.

We observed similar results for the 20 Hz limited-resource model (**Figure 4.15**). There were significant main effects of target load and speed condition on model tracking performance (target load: $F(5,45)=555.6$, $p<0.001$ for simulated group1, $F(5,45)=869.5$, $p<0.001$ for simulated group2; speed : $F(3,27)=46.3$, $p<0.001$ for simulated group1, $F(3,27)=321.4$, $p<0.001$ for simulated group2). The RMSE was marginal to significantly larger for the limited-resource model than the resource-free model (group1: 0.186 vs. 0.168, $t(9)=2.04$, $p=0.07$; group2: 0.188 vs. 0.165, $t(9)=2.66$, $p=0.026$), suggesting that the limited-resource model deviated more from the real human performance.

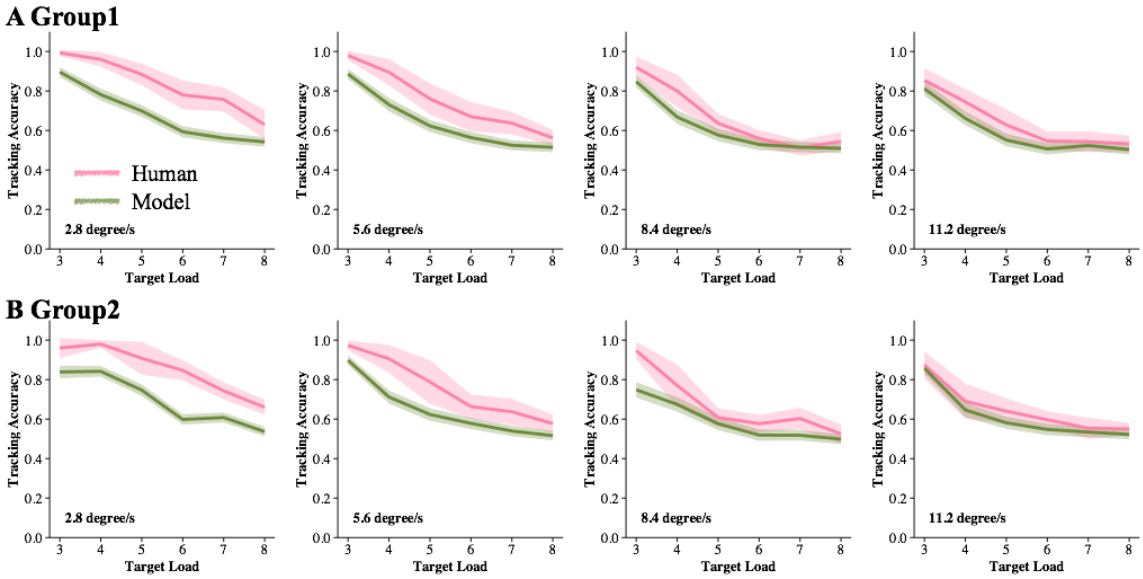


Figure 4.15 Average group 1 human participants (10 participants did different trials, panel **A**) and group 2 human participants (10 participants did the same 120 trials, panel **B**) tracking performance, together with simulated 12 Hz limited-resource model tracking performance as a function of target load and speed.

We also did a partial correlation analysis between human and model performance at the individual observer level (**Figure 4.16**). After controlling the effect of target load and speed, there was a significant correlation between the performance of human and the

limited-resource model for group1 (12 Hz: r(236)=0.447, 20 Hz: r(236)=0.437, ps<0.001) and for group2 (12 Hz: r(236)=0.427, 20 Hz: 0.459, ps<0.001). We directly compared the correlation coefficients of the resource-free and limited-resource model. The correlation coefficients of the limited-resource model did not significantly different from that of the resource-free model (all ps >0.5).

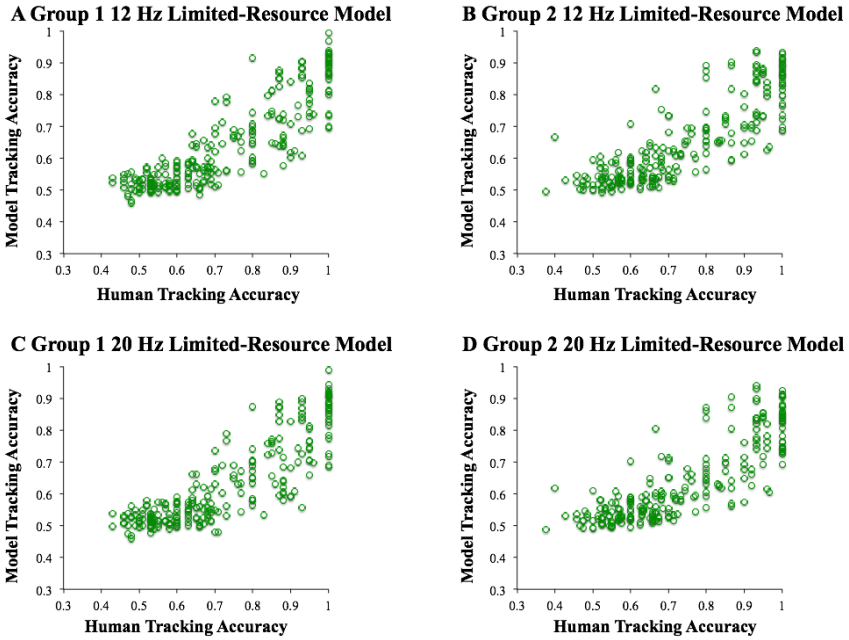


Figure 4.16 Correlations between human and limited-resource model performance across different individuals. Each dot represents a participant at a particular target load and speed condition.

Finally, we also did the analysis to test whether the limited-resource model would make similar responses as human observers. For group2 participants, there were significant positive correlations between human and model selection probability for both targets (12 Hz: r(658)=0.519, 20 Hz: r(658)= 0.522, ps<0.001) and distractors (12 Hz: r(658)=0.468, 20 Hz: r(658)=0.496 ,p<0.001, **Figure 4.17**). For the targets, the correlations we observed from the resource-free model did not significantly different

from the limited-resource model (z=1.38, p=0.17). However, for the distractors, the correlation coefficients for the limited-resource model was significantly larger than that from the resource-free model (z=2.12, p=0.03).

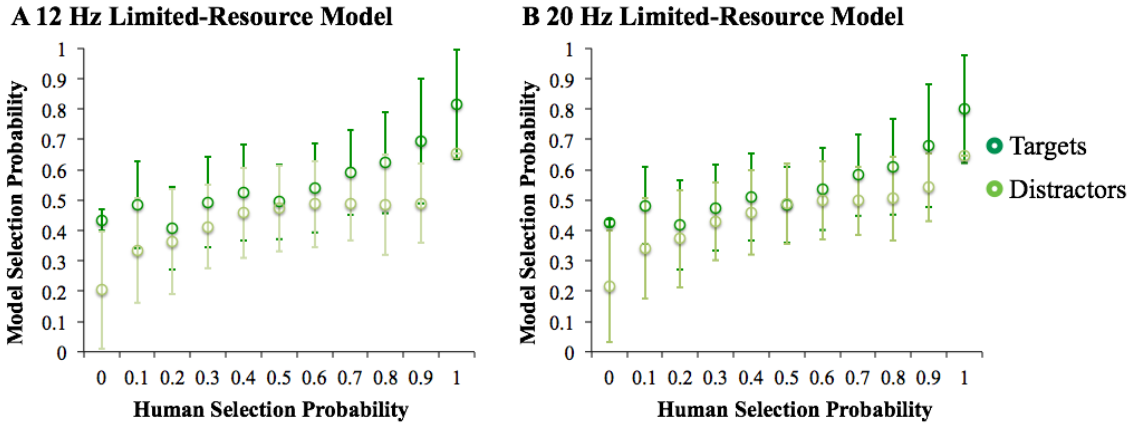


Figure 4.17 Correlation between human and the limited-resource model behavior in selecting a certain object as target. The x axis shows the probability that human observers have selected an object as target, the y axis shows the probability range that the model have selected the same object as target. Dark and light green colors represent the selection probability for real targets and real distractors respectively.

4.5.2 Discussion

In this experiment, I varied the spatial resolution at fovea to bring the constraints from limited cognitive resource into our model. In most our analysis, the limited-resource model could not explain more variance in human performance than the previous introduced resource-free model. Moreover, when measured by the general similarity to human tracking performance, adding a form of limited resource sometimes decreased the goodness of fit of the model. In sum, our results suggested that external resource constrain is not necessary to explain people's decreased tracking accuracy with higher speed and larger target load. Furthermore, no resource constrain is needed to capture the

performance variance among different individual observers and among different objects within the same tracking trial.

4.6 Computational Experiment 11: Confirming the influence of eye-movement pattern to MOT performance with the computational model

In Experiment 9, I've shown that eye-movement pattern seems to play a very important role in determining people's tracking performance. Then in Experiment 10, we've further shown that a model took real human eye-movement patterns could capture individual differences in tracking performance very well. Therefore, it seems clear that eye-movement is an important factor that constrains human tracking performance. However, many questions still need to be answered. For example, is there an optimal eye-movement pattern to maximize tracking performance for each trial? If so, how will a person, or the model, perform given this set of optimal eye-movement pattern? Will we still see target load and speed effect in performance with this set of eye-movement pattern? To address these questions, we first ran a maximum likelihood algorithm to figure out the optimal eye gaze locations at each frame in each trial. Then, we ran the model described in Experiment 10 with these eye-movement patterns, and directly compared human and model performance. As a preview, we found that this model served as an upper limit to human performance. More importantly, similarly to human observers, the model that used the optimal eye gaze locations also performed worse with higher target load and speed, further confirming the idea that the limits in tracking performance arise naturally from the limited temporal resolution, spatial resolution, and probabilistic computation.

4.6.1 Finding the optimal eye gaze locations using a maximum likelihood algorithm

It has been shown that the uncertainty about object locations, and confusions between targets and nontargets, are the primary cause of tracking errors (Bae & Flombaum, 2012). If people are representing each target and nontarget with a two dimensional Gaussian distribution, then the more overlapping between the distributions of a target and a distractor, the more likely the two would be confused and thus a tracking error would happen. Therefore, for each tracking frame, the optimal eye gaze location should be the one that can minimize the sum of overlapping areas between all target-nontarget pairs. In this case, the overall probability of confusing targets and distractors will be minimized. Given that the standard deviation of each two dimensional Gaussian distribution is dependent on its center's distance to the current eye-gaze location, we need to calculate the overlapping areas for each potential eye-gaze locations, and obtain the one that can minimize the sum of target-distractor overlaps.

At a specific tracking frame, for one potential eye gaze location (a, b) and one specific target i, the target distribution is given by

$$N\left(\mu_i, \Sigma_{i;a,b}\right) \tag{10}$$

where μ_i denotes the coordinates of the target location (x_i, y_i), and $\Sigma_{i;a,b}$ is the covariance matrix of the two dimensional distribution $\begin{bmatrix} \sigma_{i;a,b}^2 & 0 \\ 0 & \sigma_{i;a,b}^2 \end{bmatrix}$. From Experiment 8, we know that the value of $\sigma_{i;a,b}$ is dependent on the target's distance to the current eye gaze location (a, b) and could be calculated by

$$\sigma_{i;a,b} = 0.08(1 + 0.42D_{i;a,b}) \tag{11}$$

where $D_{i;\,a,\,b}$ is the distance between the target location and the eye gaze location such that

$$D_{i;a,b} = \sqrt{(x_i - a)^2 + (y_i - b)^2} \tag{12}$$

Similarly, for one specific distractor j, the distribution is given by

$$N\,(\mu_j, \Sigma_{j;a,b}) \tag{13}$$

and its mean and variance could be calculated in the same way as for the target.

The overlapping area between a target distribution and a distractor distribution could be calculated by getting the smaller value of the two distribution at all points in the space and summing these values together, and thus is given by

$$O(i,j) = \iint_{-\infty}^{+\infty} \min\left(N\left(m,n;\ \mu_i, \Sigma_{i;a,b}\right), N\left(m,n;\ \mu_j, \Sigma_{j;a,b}\right)\right) dmdn \tag{14}$$

where m and n denote the coordinates of all possible spatial locations in the visual display, and $N\left(m,n;\ \mu_i, \Sigma_{i;a,b}\right)$ denotes the probability density of target i's distribution evaluated at a specific coordinate m and n. In real calculation, we could approximate this by using summations over a finite number of discrete m and n values of m and n to simplify the calculation.

Equation (14) gives us an estimate of the amount of overlap between one target and one distractor distributions given a specific eye gaze location (a, b). To get the optimal eye gaze location at this frame, we need to find the (a, b) pair that can minimize the sum of all target and distractor overlaps. This process could be formalized by the following equation:

$$(a,b)_{optimal} = \underset{a,\,b}{argmin} \sum_i \sum_j O(i,j) \tag{15}$$

We applied equation 15 to each tracking frame in each of the 120 MOT trials done by group 2 participants in Experiment 10. Since this will require a huge amount of computations, we searched every five pixels instead of every pixel in the display. The outputs of this algorithm were 120 sets of eye-movement patterns; each contained 600 eye gaze locations, one for each frame of that trial. These eye gaze locations are supposed to be the optimal locations to look at during each specific trial.

4.6.2 Model simulation and result

To evaluate whether the obtained eye-movement patterns are really optimal, we ran our resource-free model (as described in section 4.4.1) with these sets of eye gaze locations. All parameters were the same as described in section 4.4.1, except that this time the model did not use real human eye-movement patterns.

The tracking performance of the optimal eye-movement model was plotted in **Figure 4.18**, together with corresponding human tracking performance. It's obvious from the figure that the model in general performed better than human participants. However, model also performed worse as target load or speed got higher. These intuitions were confirmed by statistical analysis. Since this time all 100 times simulation were run with the same set of eye-movement pattern, we didn't have 10 simulated participants that were comparable to human participants. Therefore, we treated each trial as one "subject" in the following analysis, and see whether the model shared any similarity to human participants at individual trial level.

We first ran paired-wise t-tests between model and human performance. The results showed that both the 12 Hz model ($t(119)=7.25$, $p<0.001$) and the 20 Hz model ($t(119)=8.87$, $p<0.001$) had significantly higher tracking accuracy than human

participants. We then ran two 6(target load) * 4(speed) univariate ANOVAs on the 12 Hz model and 20 Hz model performance. For both models, there were significant main effects of target load (12 Hz: $F(5, 96)=51.3$, $p<0.001$; 20 Hz: $F(5, 96=54.2$, $p<0.001$)), speed (12 Hz: $F(3, 96)=24.7$, $p<0.001$; 20 Hz: $F(3, 96)=10.6$, $p<0.001$), and significant interaction between the two factors (12 Hz: $F(15, 96)=1.8$, $p=0.045$; 20 Hz: $F(15, 96)=1.9$, $p=0.032$). In sum, although the model with the optimal eye gaze locations performed better than human participants, it still showed the typical target load and speed effect in tracking performance.

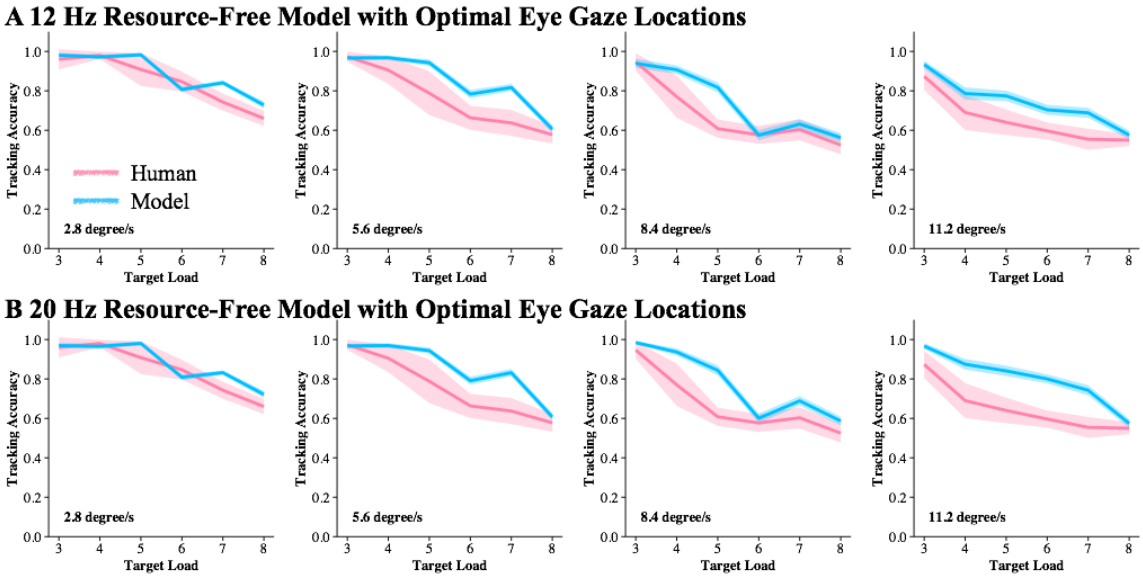


Figure 4.18. Average tracking performance of the resource-free model that used optimal eye gaze locations found by our maximum likelihood algorithm, plotted together with human performance from group2 participants. Both the 12 Hz and 20 Hz model performed much better than human observers.

4.6.3 Discussion

In this experiment, I first used a maximum likelihood algorithm to find the optimal eye gaze locations at each frame of an MOT trial. We then ran our resource-free

model with these sets of optimal eye-movement patterns and found that this model's performance could serve as an upper limit of human tracking performance. More importantly, even with the optimal eye-gaze locations, the model still performed worse with more targets and higher speed.

At the first look of these results, one may feel surprised that a model with a set of optimal eye-gaze locations would still make many errors at higher target load and speed conditions. However, similar to the model that adopted human eye-movement patterns, the current model is still constrained by a number of factors that can affect tracking performance, i.e. the imperfect temporal resolution, spatial resolution, and probabilistic correspondence computation. Similar to the idea under ideal observer models, the current model is only optimal given available information and specific constrains (Geisler, 2011). The current results suggested that even eye-movements is optimal, tracking errors are still *unavoidable* given the constraints imposed by the human visual system, especially at higher target load and speed conditions. Although eye-movement selection is very important in determining tracking performance, it alone cannot completely explain the limits of human tracking. The imperfect nature of human spatial and temporal resolution, together with the probabilistic computation, are also critical factors that are limiting human tracking ability. The presence of the target load effect and speed effect in the performance of optimal eye-movement model further supported the idea that external resource limits are not necessary to explain the capability and limit of human tracking.

Finally, it's also worth noting that the current optimal eye-movement patterns were calculated in a static way. That is, we treated each tracking frame as independent from each other, and there is no cost of making too many saccades. Future studies should

try to incorporate these factors and find optimal eye-movement strategies that more similar to the eye-movements used by real human observers.

4.7 General Discussion

In this study, I have shown that both constraints on visual inputs and computational algorithms are critical factors liming human tracking ability. I first used two behavioral experiments to measure the temporal and spatial limits of the visual inputs of a typical MOT task. I then used an eye-tracking experiment to show that eye gaze locations have a great influence on tracking performance by determining the spatial resolution of different objects in the display. Finally, I built a computational model of MOT that relies on probabilistic correspondence computations. This model is constrained by the temporal and spatial resolutions I measured in Experiment 7 and 8. It is also constrained by real human eye-movements I collected from 20 participants. Without fitting any parameter, the model performed similarly to human participants. The model performed worse with higher target load and faster speed conditions. Furthermore, the model could also capture individual differences among participants, as well as difference among objects in a same trial. These results suggested that tracking limits could be clearly observed as long as constraints imposed by the human visual system were considered. Experiment 10b directly test the alternative hypothesis that a model with an external resource constraint could explain the tracking limits at higher target load and speed conditions better (Alvarez & Franconeri, 2007; Ma & Huang, 2009; Vul et al, 2009; Holcombe & Chen, 2011). However, the results suggested that adding a form of external resource limit could not bring any marginal benefit in terms of fitting human

data. What's more, the limited-resource model significantly underestimate human performance, and the goodness-of-fit is significantly worse than that of the resource-free model. Finally, in Experiment 11, I combined a maximum likelihood algorithm and our resource model to show that for each MOT trial, there did exist a set of optimal eye-movement pattern that can make the model do better than using other eye-movement patterns. However, even this model has decreased performance at higher target load and speed conditions. This result further supported the idea that external resource constraint is not necessary to explain human tracking limits. It is the inherent perceptual limit, eye-movement selection, and probabilistic computation that limit human tracking ability.

Visual-cognitive limitations are usually thought to reflect the consumption of limited neural commodities in the brain, that is, limitations imposed on the amount of memory and/or attention available to expend in a task (Vul et al., 2009; Holcombe & Chen, 2013; Luck & Vogel, 2013; Ma, Husain, & Bays, 2014). Contrary to the general assumption in the literature, I found that a model that considering eccentricity dependent noise, limited temporal resolution human observers' eye-movement strategies, and most importantly the noisy correspondence computations could produce the typical target load and speed effects taken as evidence of cognitive resources. These results suggest that at least tracking abilities my be limited entirely from the 'bottom-up,' through the computational demands of the task interacting with the non-uniform quality of visual inputs obtained by the human retina.

The current resource-free model couldn't explain all of the variability in human performance. We suspected that at least two other factors should be considered in future research. First, given the probabilistic nature of visual tracking, there is a lot of inherent

noise (independent from target load, speed, or any other trajectory factors) in the human tracking system. Different observers will show different behaviors to the same trial, and even the same observer will show different behaviors when complete the same trial twice. In a previous paper, I have shown that the split-half correlation among human tracking performance is about 0.56. Therefore, at the first place, there is not much the to-be-explained variance. Second, it has been shown that human observers differ a lot in terms of their spatial and temporal resolutions. Since we didn't put these values as free parameters in the model, we weren't able to capture the entire individual difference space. However, without separately fitting these values, we've already seen a good model fit to individual difference. I expect that more studies on computational modeling of visual tracking, taking the inherent limit of visual perception as well as implementations of the correspondence computation process, would finally explain the mystery of the behavioral limits of human visual tracking.

Chapter 5: General discussion and conclusion

In three studies, I have shown that how correspondence computation could be done during some classic visual tasks. I've also shown that how errors in correspondence computations could lead to observed limits in human behavior.

In the first study, I've shown that corresponding noisy signals to their source and forming individuated object representation is a challenging task for human observers. A density-based clustering algorithm considering eccentricity-dependent noise could simulate human behavior very well. These results suggested that correspondence computation is playing an important role even at the start of a working memory task. At least part of the observed limit in memory performance should be due to imperfect correspondence computation to individual objects.

In the second study, I've shown how different correspondence analyses are combined to generate coherent motion perceptions. Both the lower-level transient detection system and higher-level position comparison system are playing important roles. The relative contributions of each system depend on the strength of signals they can get.

In the last study, I've combined behavioral data and computational models to show how correspondence computations, under the constraints from physical limits, could support object-tracking ability. Inherent limits in tracking will arise if noisy, discrete visual inputs and probabilistic correspondence computations are considered. No external commodity-like cognitive resources is needed to simulate human performance. Moreover, part of individual differences in tracking performance could be explained by different eye-movement strategies.

Overall, these results suggest that correspondence computation is a pervasive process that serves as the primary constraint on many visual cognitive tasks. So why does its role is largely underestimated in previous literature? Presumably because people can often complete correspondence computations effortlessly even without awareness. In addition to this, too much emphasis on the resource-based limits has made researchers ignore the real computations and algorithms happening in the human mind.

Marr (1982) has proposed that an information system must be understood at three levels. The first level is the computation theory level, which specifies the goal of a specific computation and describes its appropriateness for the task. The second level is the representation and algorithm level, which needs to decide the representation for the input and output of the system, as well as the algorithm used to do transformations between theses representations. The third level is the implementation level, which needs to determine how the computations are implemented physically with some hardware. The three levels are relatively independent, but could affect each other given their own properties and constraints.

It's true that some limits of an information system could be explained by resource limits in the implementation level. However, problems that arise at the computational and representational levels couldn't be fully explained by only looking at how they are implemented. As discussed in Chapter 4, the fact that Bubble sort is much slower than Quicksort couldn't be explained by limited memory resource of the computer. Rather, the difference in speed is easily explained if one directly compared the difference in the two algorithms. Focusing on limited resource is misleading in this case, and will let people make totally wrong conclusions.

Therefore, it's very important for vision scientists to shift to a more computational and algorithm-based approach to study the human mind. Since correspondence computation is supporting many human cognitive abilities, studying the mechanisms and roles of correspondence computations should provide a more unified explanation to many phenomena we observed in human behavior. This approach will also help us get rid of the homonculus problem. Since this approach emphasis on making every representation and computation explicit, we should naturally observe the cognitive abilities and limits, without assuming a limited-resource homonculus in the brain who is controlling how the brain would complete the tasks.

References

Adelson, E. H., & Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature, 300,* 523-525. doi:10.1038/300523a0

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *JOSA, 2,* 284-299. doi: 10.1364/JOSAA.2.000284

Alvarez, G. & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106-111.

Alvarez, G., & Franconeri, S. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7(13):* 14, 1–10.

Anderson, D. E., & Awh, E. (2012). The plateau in mnemonic resolution across large set sizes indicates discrete resource limits in visual working memory. *Attention, Perception, and Psychophysics, 74,* 891– 910.

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a plateau when individual item-limits are exceeded. *Journal of Neuroscience, 31*, 1128-1138.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in cognitive sciences,16(8),* 437-443.

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & psychophysics, 55(5),* 485-496.

Bacon, W. F., & Egeth, H. E. (1997). Goal-directed guidance of attention: evidence from conjunctive visual search. *Journal of Experimental Psychology: Human Perception and Performance, 23(4),* 948.

Bae, G. Y., & Flombaum, J. I. (2012). Close encounters of the distracting kind: Identifying the cause of visual tracking errors. *Attention, Perception, & Psychophysics, 74(4),* 703-715.

Bae, G. Y., & Flombaum, J. I. (2013). Two items remembered as precisely as one: How integral features can improve visual working memory. *Psychological Science, 24,* 2038-2047.

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744-763.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision,9(12),* 13.

BarShalom, Y., Daum, F., & Huang, J. (2009). The probabilistic data association filter. *Control Systems, IEEE, 29(6),* 82–100.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86(3),* 201-221.

Battelli, L., Cavanagh, P., Intriligator, J., Tramo, M. J., Hénaff, M., Michèl, F., & Barton, J. S. (2001). Unilateral right parietal damage leads to bilateral deficit for high-level motion. *Neuron, 32,* 985-995. doi:10.1016/S0896-6273(01)00536-0

Battelli L., Pascual-Leone A., & Cavanagh, P. (2007). The 'when' pathway of the right parietal lobe. *Trends in Cognitive Sciences, 11,* 204-210.

doi:10.1016/j.tics.2007.03.001

Baylis, G. C., & Baylis, L. L. (2001). Visually misguided reaching in Balint's syndrome. *Neuropsychologia, 39,* 865-875. doi:10.1016/S0028-3932(01)00009-4

Baynes K., Holtzman, J. D. & Volpe, B. T. (1986). Components of visual attention: Alterations in response pattern to visual stimuli following parietal lobe infarction. *Brain, 109,* 99-114. doi: 10.1093/brain/109.1.99

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9(10):*7, 1–11,

Bays, P.M. & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*, 851-854.

Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 1, p. 740). New York: Springer.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*, 177–178.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research, 14,* 519-527. doi: 10.1016/0042-6989(74)90041-8

Braddick, O. J., & Adlard, A. (1978). Apparent motion and the motion detector. *In Visual Psychophysics and Physiology,* ed. J. C. Armington, J. Krauskopf, B. R. Wooten. New York: Academic.

Brady, T. F. and Alvarez, G.A. (2011). Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychological Science*, *22*, 384-392.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working

memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review, 120(1),* 85.

Brainard, D. (1997). The psychophysics toolbox. Spatial Vision, 10, 433–436.

Brannon, E. M. (2006). The representation of numerical magnitude. *Cur- rent Opinion in Neurobiology, 16,* 222–229. doi:10.1016/j.conb.2006.03 .002

Breitmeyer, B. G. (1984). Visual Masking: An integrative approach. Oxford University Press.

Breitmeyer, B. G., & Ritter, A. (1986). Visual persistence and the effect of eccentric viewing, element size, and frame duration on bistable stroboscopic motion percepts. *Perception & Psychophysics, 39(4),* 275-280.

Britten, K. H., & Heuer, H. W. (1999). Spatial summation in the receptive fields of MT neurons. *The Journal of Neuroscience, 19,* 5074-5078.

Burke, L. (1952). On the tunnel effect. *Quarterly Journal of Experimental Psychology, 4,* 121-138. doi: 10.1080/17470215208416611

Burr, D., & Thompson, P. (2011). Motion psychophysics: 1985–2010. *Vision Research, 51,* 1431-1456. doi:10.1016/j.visres.2011.02.008

Burr, D., & Ross, J. (2008). A visual sense of number. *Current Biology, 18,* 425–428.

Carrasco, M., & Frieder, K. S. (1997). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research*, *37*(1), 63-82.

Cavanagh, P. (1992). Attention-based motion perception. *Science, 257,* 1563-1565. doi: 10.1126/science.1523411

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9,* 349–354.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 109-127.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in cognitive sciences, 15(8),* 358-364.

Colas, F., Flacher, F., Tanner, T., Bessiere, P., & Girard, B. (2009). Bayesian models of eye movement selection with retinotopic maps. *Biological Cybernetics*, *100*(3), 203-214.

Cox, D., D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience, 8(9),* 1145-1147.

Daniel, P. M., & Whitteridge, W. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of physiology*, *159*(2), 203-221.

Dawson, M. R. W. (1991). The how and why of what went where in apparent motion: modeling solutions to the motion correspondence problem. *Psychological Review, 98,* 569. doi: 10.1037/0033-295X.98.4.569

Dawson, M. R. W., & Wright, R. D. (1994). Simultaneity in the Ternus configuration: Psychophysical data and a computer model. *Vision Research, 34,* 397-407.

Dayan, P., & Solomon, J. A. (2010). Selective Bayes: Attentional load and crowding. *Vision research, 50(22),* 2248-2260.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11(8),* 333-341.

Di Lollo, V., Kawahara, J. I., Ghorashi, S. S., & Enns, J. T. (2005). The attentional blink: resource depletion or temporary loss of control? *Psychological research, 69(3),* 191-200.

di Pellegrino, G., & de Renzi, E. (1995). An experimental investigation on the nature of extinction. *Neuropsychologia, 33,* 153-170. doi:10.1016/0028-3932(94)00111-2.

Drasdo, N., Fowler, C. W. (1974). Non-linear projection of the retinal image in a wide-angle schematic eye. *The British Journal of ophthalmology,* 58, 709-714.

Drew, T., & Vogel, E. K. (2008). Neural measures of individual differences in selecting and tracking multiple moving objects. *The Journal of Neuroscience,28(16),* 4183-4191.

Dux, P. E., & Marois, R. (2009). The attentional blink: A review of data and theory. *Attention, Perception, & Psychophysics, 71(8),* 1683-1700.

Egeth, H., & Dagenbach, D. (1991). Parallel versus serial processing in visual search: further evidence from subadditive effects of visual quality. *Journal of Experimental Psychology: Human Perception and Performance, 17(2),* 551.

Enns, J. T., & Di Lollo, V. (1997). Object substitution: A new form of masking in unattended visual locations. *Psychological Science, 8(2),* 135-139.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in cognitive sciences, 4(9),* 345-352.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD, 96(34),* 226-231.

Feldman, J., Singh, M., & Froyen, V. (in press) Perceptual grouping as Bayesian mixture estimation. Gepshein, S., Maloney, L. & Sign, M. (Eds.) *Oxford Handbook of Computational Perceptual Organization*.

Fehd, H. M., & Seiffert, A. E. (2008). Eye movements during multiple object tracking: Where do participants look?. *Cognition*, *108*(1), 201-209.

Fehd, H. M., & Seiffert, A. E. (2010). Looking at the center of the targets helps multiple object tracking. *Journal of vision*, *10*(4), 19.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8,* 307–314.

Franconeri, S. L., Jonathan, S., & Scimeca, J. (2010). Tracking multiple objects is limited only by object spacing, not speed, time, or capacity. *Psychological Science, 21,* 920–925.

Friedman-Hill, S. R., Robertson, L. C., & Treisman, A. (1995). Parietal contributions to visual feature binding: Evidence from a patient with bilateral lesions. *Science, 269,* 853-855.

Fobes, J. L., & King, J. E. (1982). Primate behavior. New York: Academic Press.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working mem- ory. *Nature Communications, 3,* 1229

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research.*Vision research*, *51*(7), 771-781.

Goldstein, E. B., & Fink, S. I. (1981). Selective attention in vision: Recognition memory for superimposed line drawings. *Journal of Experimental Psychology: Human Perception and Performance, 7(5),* 954-967.

Goss, R. G., & Nitschke, G. S. (2014). Automating Network Protocol Identification. Issac, B., & Israr, N., (Eds.) *Case Studies in Intelligent Computing: Achievements and Trends*. CRC press.

Grandison, T. D., Ghirardelli, T. G., & Egeth, H. E. (1997). Beyond similarity: Masking of the target is sufficient to cause the attentional blink. *Perception & Psychophysics, 59(2),* 266-274.

Haladjian, H. H., & Pylyshyn, Z. W. (2011). Enumerating by pointing to locations: A new method for measuring the numerosity of visual object representations. *Attention, Perception, & Psychophysics, 73,* 303–308.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology, 44(5),* 1457.

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition*, *7*(1-3), 43-64.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383,* 334–337. doi:10.1038/383334a0

He, Z. J., & Ooi, T. L. (1999). Perceptual organization of apparent motion in the Ternus display. Perception, 28(7), 877-892.

Hildreth, E. C. (1984). Computations underlying the measurement of visual motion. Artificial Intelligence, 23, 309-354. doi:10.1016/0004-3702(84)90018-3

Hock, H. S., Gilroy, L., & Harnett, G. (2002). Counter-changing luminance: A non-Fourier, nonattentional basis for the perception of single-element apparent motion.

Journal of Experimental Psychology: Human Perception and Performance, 28, 93-112. doi: 10.1037//0096-1523.28.1.93

Holcombe, A. O., Chen, W. (2011). Exhausting attentional tracking resources with a single fast-moving object*. Cognition, 123,* 218-228.

Holcombe, A. O., & Chen, W. Y. (2013). Splitting attention reduces temporal resolution from 7 Hz for tracking one object to< 3 Hz when tracking three.*Journal of Vision*, *13*(1), 12.

Horowitz, T., & Cohen, M. (2010). Direction information in multiple object tracking is limited by a graded resource. *Attention, Perception, & Psychophysics, 72,* 1765-1775.

Humphreys, G. W., & Riddoch, M. J. (1993). Interactions between object and space systems revealed through neuropsychology. In D. E. Meyer, & S. Kornblum (Eds), *Attention and performance XIV* (pp. 143-162). Cambridge, MA: MIT Press.

Im, H. Y., Zhong, S. H., & Halberda, J. (in press). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision research*.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. Cognitive Psycholo- gy, 43, 171–216.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line.*Cognition*, *106*(3), 1221-1247.

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 683-702.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of the object files:

Object-specific integration of information. *Cognitive Psychology, 24,* 174–219.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82(1),* 35–45.

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature Reviews Neuroscience, 1,* 91-100.

Kramer, P., & Yantis, S. (1997). Perceptual grouping in space and time: Evidence from the Ternus display. *Perception & Psychophysics, 59(1),* 87-99.

Keshvari, S., van den Berg, R., & Ma, W. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology, 9(2),* e1002927

Kouider, S., De Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in cognitive sciences, 14(7),* 301-307.

Lages, M. (2013). Straight or curved? From deterministic to probabilistic models of 3D motion perception. *Frontiers in Behavioral Neuroscience, 7(79),* 1-3. doi: 10.3389/fnbeh.2013.00079

Lages, M. & Heron, S. (2010). On the inverse problem of binocular 3D motion perception. *Plos computational biology, 6*(11), e1000999. doi: 10.1371/journal.pcbi.1000999

Lamy, D., Leber, A., & Egeth, H. E. (2004). Effects of task relevance and stimulus-driven salience in feature-search mode. *Journal of Experimental Psychology: Human Perception and Performance, 30(6),* 1019

Lamme, V. A. F. (2003). Why visual attention and awareness are different. Trends in Cognitive Sciences, 7, 12-18.

Landau, A. N., & Fries, P. (2012). Attention samples stimuli rhythmically. *Current Biology, 22(11),* 1000– 1004.

Leber, A. B., & Egeth, H. E. (2006). It's under control: Top-down search strategies can override attentional capture. *Psychonomic Bulletin & Review,13(1),* 132-138.

Levi, D. M. (2008). Crowding – An essential bottleneck for object recognition: A mini-review. *Vision Research, 48,* 635-654.

Levi, D. M., Hariharan, S., & Klein, S. A. (2002). Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of Vision, 2*(2), 3. doi:10.1167/2.2.3

Li, J., Shao, N., Xu, H., Shui, R., Shen, M. (2013). Does visual working memory work as a few fixed slots? *The Quarterly Journal of Experimental Psychology, 66(11),* 2103-2117.

Liu, G., Austen, E. L., Booth, K. S., Fisher, B. D., Argue, R., Rempel, M. I., Enns, J. T. (2005). Multiple object tracking is based on scene, not retinal, coordinates. *Journal of Experimental Psychology: Human Perception & Performance, 31,* 235–247.

Lu, Z. L., & Sperling, G. (1996). Three systems for visual motion perception. *Current Directions in Psychological Science, 5(2),* 44-53.

Lu, Z. L., & Sperling, G. (2001). Three-systems theory of human visual motion perception: review and update. *JOSA A, 18(9),* 2331–2370. doi: 10.1364/JOSAA.18.002331

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences in cognitive ability. *Trends in Cognitive Sciences, 17,* 391-400

Ma, Z., & Flombaum, J. I. (2013). Off to a bad start: Uncertainty about the number of targets at the onset of multiple object tracking. *Journal of experimental psychology: human perception and performance, 39(5),* 1421-1432.

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision, 9(11),* 3, 1-30.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17(3),* 347-356.

Mack, A., & Rock, I. (1998). Inattentional blindness. Cambridge, MA: MIT press.

Makovski, T., Vázquez, G. A., & Jiang, Y. V. (2008). Visual learning in multiple-object tracking. *Plos One,* e2228.

Marr, D. (1982). Vision. New York: Freeman.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review, 63(2),* 81-97.

Murphy, K. P. (2012). Machine learning: A probabilistic perspective. Cambridge, MA: The MIT Press.

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320(6059),* 264-265.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive psychology, 7(4),* 480-494.

Neri, P., & Levi, D. M. (2006). Spatial resolution for feature binding is impaired in peripheral and amblyopic vision. *Journal of Neurophysiology, 96(1),* 142-153.

Ogawa, H., Watanabe, K., & Yagi, A. (2009). Contextual cueing in multiple object tracking. *Visual Cognition*, *17*(8), 1244-1258.

Oksama, L., & Hyona, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition, 11,* 631-671.

Oksama, L., & Hyona, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology, 56,* 237–283.

Orhan, A. E., & Jacobs, R. A. (2014a). Toward ecologically realistic theories in visual short-term memory research. *Attention, Perception, & Psychophysics,76(7),* 2158-2170.

Orhan, A. E., & Jacobs, R. A. (2014b). Are performance limitations in visual short-term memory tasks due to capacity limitations or model mismatch? *arXiv preprint arXiv:1407.0644*.

Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The adaptive nature of visual working memory. *Current Directions in Psychological Science,23(3),* 164-170.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437–442.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4(12),* 12. doi:10.1167/4.12.12

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11,* 1129-1135. doi:10.1038/nn.2187

Petersik, J. T. (1989). The two-process distinction in apparent motion. *Psychological Bulletin, 106,* 107-127. doi: 10.1037/0033-2909.106.1.107

Petersik, J. T. (1995). A comparison of varieties of "second-order" motion. *Vision Research, 35,* 507-517. doi:10.1016/0042-6989(94)E0092-Y

Petersik J. T. (2006). The evolution of explanations of a perceptual phenomenon: A case history using the Ternus effect. *Perception, 35,* 807-821.

Petersik J. T. (2009). Orientation anisotropy in the Ternus phenomenon. *Perceptual and Motor Skills, 108,* 405-410.

Petersik, J. T., & Pantle, A. (1979). Factors controlling the competing sensations produced by a bistable stroboscopic motion display. *Vision Research, 19(2),* 143-154.

Pierce, R. S., Bian, Z., Braunstein, M. L., & Anderson, G. (2013). Detection of 3D curved trajectories: the role of binocular disparity. *Frontiers in Behavioral Neuroscience, 7(12),* 1-6. doi: 10.3389/fnbeh.2013.00012

Platt, J. R., & Johnson, D. M. (1971). Localization of position within a homogeneous behavior chain: Effects of error contingencies. *Learning and Motivation, 2,* 386–414. doi:10.1016/0023-9690(71)90020-8

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial index model. *Cognition, 32,* 65-97.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. Cognition, 80, 127-158.

Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking (MOT): I. Tracking without keeping track of object identities. *Visual Cognition, 11,* 801-822.

Pylyshyn, Z. W., & Annan, V. (2006). Dynamics of target selection in multiple object tracking (MOT). Spatial Vision, 19, 485-504.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 179-197.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849-860.

Ramachandran, V. S., & Anstis, S. M. (1986). The perception of apparent motion. *Scientific American, 254(6),* 102-109. doi: 10.1038/scientificamerican0686-102

Reichardt, W. (1961). Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In W. A. Rosenblith (Ed), *Sensory Communication,* (pp. 303-317). Cambridge, MA: MIT Press

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological science*, *8*(5), 368-373.

Robertson, L., Treisman, A., Friedman-Hill, S., & Grabowecky, M. (1997). The interaction of spatial and object pathways: Evidence from Balint's syndrome. *Journal of Cognitive Neuroscience, 9,* 295-317. doi: 10.1162/jocn.1997.9.3.295

Rossetti, Y., Pisella, L., & Vighetto, A. (2003). Optic ataxia revisited. *Experimental Brain Research, 153(2),* 171-179. doi: 10.1007/s00221-003-1590-6

Rovamo,J. & Virsu, V. (1979). An estimation and application of the human cortical magnificationfactor. *Experimental Brain Research, 37, 495–510.*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., & Berg, A. C. (2015). Imagenet large scale visual cognition challenge. *International Journal of Computer Vision, 115(3),* 211-252.

Scholl, B. J. (2009). What have we learned about attention from multiple object tracking (and vice versa). In D. Dedrick & L. Trick (Eds.), Computation, Cognition, and Pylyshyn (pp. 49–78). Cambridge, MA: MIT Press.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology, 38,* 259–290.

Sears, C. R., & Pylyshyn, Z. W. (2000). Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology, 54,* 1-14.

Seiffert, A. E., & Cavanagh, P. (1998). Position displacement, not velocity, is the cue to motion detection of second-order stimuli. *Vision Research, 38,* 3569-3582. doi: 10.1016/S0042-6989(98)00035-2

Seiffert, A. E., & Cavanagh, P. (1999). Position-based motion perception for color and texture stimuli: effects of contrast and speed. *Vision Research, 39,* 4172-4185. doi: 10.1016/S0042-6989(99)00129-7

Shapiro, K., Driver, J., Ward, R., & Sorensen, R. E. (1997). Priming from the attentional blink: A failure to extract visual tokens but not visual types. *Psychological Science*, *8*, 95-100.

Shimojo, S., Silverman, G. H., Nakayama, K. (1989). Occlusion and the solution to the aperture problem for motion. *Vision Research, 29,* 619-626. doi: 10.1016/0042-6989(89)90047-3

Simons, D. J. (2000). Current approaches to change blindness. *Visual cognition, 7(1-3),* 1-15.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London, 28(9),* 1059-1074.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in cognitive sciences, 1(7),* 261-267.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in cognitive sciences, 9(1),* 16-20.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review, 119(4),* 807-830.

Spelke, E. S., Kestenbaum, R., Simons, D. J., & Wein, D. (1995). Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology, 13(2),* 113-142.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74,* 1–29.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51(6),* 599-606.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta Psychologica, 135(2),* 77-99.

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience, 5(2),* 144-155.

Tootell, R. B. H., Reppas, J. B., Kwong, K., K., Malach, R., Born, R. T., Brady, T. J., Rosen, B. R., & Belliveau, J. W. (1995). Functional analysis of human MT and related-visual cortical areas using magnetic resonance imaging. *Journal of Neuroscience, 15,* 3215-3230.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preatten- tive processing. *Journal of Experimental Psychology: Human Percep- tion and Performance, 19,* 331–351.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101(1),* 80-102

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology, 5(2),* 207-232.

Ullman, S. (1979). *The interpretation of visual motion.* Cambridge, MA: MIT Press.

Ullman, S. (1984). Visual routines. *Cognition, 18,* 97-159. doi: 10.1016/0010-0277(84)90023-4

van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012) Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*, 8780-8785.

van den Berg, R., Roerdink, J. B. T., M., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLoS Comput. Biol. 6*, e1000646

VanRullen, V., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences, 7(5),* 207–213.

VanRullen, R., & Macdonald, J. S. (2012). Perceptual echoes at 10 Hz in the human brain. *Current biology*, *22*(11), 995-999.

van Santen, J. P. H., & Sperling, G. (1984). Temporal covariance model of human motion perception. *JOSA, 1,* 451-473. doi: 10.1364/JOSAA.1.000451

Virsu,V. &Rovamo,J. (1979).Visualresolution,contrastsensitivity, and the cortical magnificationfactor. *ExperimentalBrain Research, 37,475-494.*

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*,*438*(7067), 500-503.

Vul, E., Frank, M. C., & Tenenbaum, J. B. (2009). Explaining human multiple objects tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems 22,* 1955–1963.

Ward, R., Duncan, J., & Shapiro, K. (1996). The slow time-course of visual attention. *Cognitive psychology, 30(1),* 79-109.

Welch, G., & Bishop, G. (2006). An introduction to the Kalman filter. An introduction to the Kalman filter. University of North Carolina: Chapel Hill, NC.

Whitney, D., & Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in cognitive sciences, 15*, 160–168.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4(12),* 11, 1-16.

Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *JOSA A, 14(9),* 2057-2068.

Wolfe, J., M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance, 15, 419-433.

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology,* 24, 295-340.Yantis & Jonides, 1990

Yi, D., Turk-Browne, N. B., Flombaum, J. I., Kim, M., Scholl, B. J., & Chun, M. M. (2008). Spatiotemporal object continuity in human ventral visual cortex. *PNAS, 105(26),* 8840-8845.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys (CSUR), 38(4),* 13, 1–45.

Young, R. W. (1971). The renewal of rod and cone outer segments in the rhesus monkey. *The Journal of cell biology*, *49*(2), 303-318.

Zelinsky, G. J., & Neider, M. B. (2008). An eye movement analysis of multiple object tracking in a realistic environment. *Visual Cognition*, *16*(5), 553-566

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233-236.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science, 22(11),* 1434-1441.

Zihl, J., von Cramon, D., & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain, 106,* 313-340. doi: http://dx.doi.org/10.1093/brain/106.2.313

Zihl, J., von Cramon, D., Mai, N, & Schmid, C. H. (1991). Disturbance of movement vision after bilateral posterior brain damage: Further evidence and follow up observations. *Brain, 114,* 2235-2252.

Zohary, E., & Hochstein, S. (1989). How serial is serial processing in vision. *Perception, 18(2),* 191-200.

Curriculum Vitae

Zheng Ma was born on May 26th, 1989 in Beijing, China. After getting a B.S. in psychology at Peking University, China, she continued to pursue a PhD degree at Johns Hopkins University. At Johns Hopkins University, she was advised by Dr. Jonathan Flombaum and had close collaborations with many other professors both inside and outside the psychology department. She has also gained a lot of teaching experience by being the TA and instructor of different classes. Her long-term goal is to become a research scientist to understand the mechanism of human vision. She is also hoping that she can make more young people get interested in psychology researches.